2013

Linking across forms in vertical scaling under the common-item nonequvalent groups design

Xuan Wang
University of Iowa

LINKING ACROSS FORMS IN VERTICAL SCALING UNDER THE COMMON-ITEM NONEQUIVALENT GROUPS DESIGN

by

Xuan Wang

An Abstract

Of a thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Psychological and Quantitative Foundations (Educational Measurement and Statistics)
in the Graduate College of
The University of Iowa

May 2013

Thesis Supervisors: Professor Donald B. Yarbrough
Adjunct Asst. Professor Deborah J. Harris

ABSTRACT

The purposes of this dissertation are to compare how different the resulting proficiency estimates are by using two scale maintenance approaches, the horizontal and vertical approaches, in supporting scale stability across grade within form, within grade across forms, and across grades and across forms, and to thus investigate under which conditions of within-grade variability patterns and examinee sample characteristics one approach is preferable to the other.

Since there is no universally accepted growth model in the literature, three different distribution sets were specified and generated with regard to within-grade variability patterns in the simulation data: constant across grades, decreasing variability as grade increases, and increasing variability as grade increases. In addition, two sets of examinee sample characteristics were also specified in the simulation data: small examinee group difference and large examinee group difference. Thus six proficiency distribution conditions were used to generate data in this dissertation. Under the six conditions of proficiency distributions, the performances of the two scale maintenance approaches on the resulting proficiency estimates across multiple forms were the foci in this dissertation.

One major implication of this study is that the accuracy in recovering the true examinee proficiencies on the new form across multiple linking appeared to be an interaction among the number of forms or years involved in the linking process, the within-grade variability patterns, and the examinee group differences, but they did not appear to be a function of the maintenance approach implemented.

The findings from this study provide important empirical guidance to practitioners on how the vertical scale can be maintained, once a vertical scale is established. If the desired output of a maintained scale is to continue to capture the characteristics of the established scale in terms of grade separation, within-grade variability, and growth implications, the vertical approach appears to be marginally better in achieving these goals. However, the differences observed across three forms are marginal, and in some cases neither approach demonstrates superiority in preserving the same patterns at the baseline scale. Both approaches are able to reasonably well capture the trend of the baseline scale, at least across three forms.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

_____
Thesis Supervisor

_____
Title and Department

_____
Date

LINKING ACROSS FORMS IN VERTICAL SCALING UNDER THE COMMON-ITEM NONEQUIVALENT GROUPS DESIGN

by

Xuan Wang

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Psychological and Quantitative Foundations (Educational Measurement and Statistics)
in the Graduate College of
The University of Iowa

May 2013

Thesis Supervisors: Professor Donald B. Yarbrough
Adjunct Asst. Professor Deborah J. Harris

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

PH.D. THESIS

_____

This is to certify that the Ph.D. thesis of

Xuan Wang

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Psychological and Quantitative Foundations (Educational Measurement and Statistics) at the May 2013 graduation.

Thesis Committee: _____
Donald B. Yarbrough, Thesis Supervisor

_____
Deborah J. Harris, Thesis Supervisor

_____
Michael J. Kolen

_____
Stephen Dunbar

_____
Liz Hollingworth

To Zhaohui

ACKNOWLEDGMENTS

I wish to express my deepest appreciation to my academic adviser and thesis supervisor Dr. Donald Yarbrough. He supported me with his cordial generosity when I was in need, and guided me with his academic insight and passion when I was lost. Without his constant guidance, advice, assistance, and encouragement, this dissertation might not been completed. It was a privilege to work with him and I cherish the memory of being his student.

My special appreciation goes to my thesis supervisor Dr. Deborah Harris. I want to thank her for her patience, her many hours mentoring me and her great insights and guidance in the creation of my dissertation and my many other works. I also appreciate her nagging me on the dissertation work throughout the years and helping me speed up the process.

I would like to express my thanks to the other committee members, Dr. Michael Kolen, Dr. Stephen Dunbar, and Dr. Liz Hollingworth, for their time and contributions to this study. Their careful review and insightful comments improved the quality of this study. I thank the Center for Evaluation and Assessment at the University of Iowa and ACT, Inc., for helping me both academically and financially in the years of my doctoral study.

I am particularly grateful to my grandmother and my parents in China, for always supporting me in countless ways and for their confidence in me. Finally and most importantly, I wish to express my endless gratitude and love to my husband Zhaohui for

his endurance, care, and indulgence in letting me pursue my own dreams. I would not have been where I am today if it were not for his continued love and support.

ABSTRACT

The purposes of this dissertation are to compare how different the resulting proficiency estimates are by using two scale maintenance approaches, the horizontal and vertical approaches, in supporting scale stability across grade within form, within grade across forms, and across grades and across forms, and to thus investigate under which conditions of within-grade variability patterns and examinee sample characteristics one approach is preferable to the other.

Since there is no universally accepted growth model in the literature, three different distribution sets were specified and generated with regard to within-grade variability patterns in the simulation data: constant across grades, decreasing variability as grade increases, and increasing variability as grade increases. In addition, two sets of examinee sample characteristics were also specified in the simulation data: small examinee group difference and large examinee group difference. Thus six proficiency distribution conditions were used to generate data in this dissertation. Under the six conditions of proficiency distributions, the performances of the two scale maintenance approaches on the resulting proficiency estimates across multiple forms were the foci in this dissertation.

One major implication of this study is that the accuracy in recovering the true examinee proficiencies on the new form across multiple linking appeared to be an interaction among the number of forms or years involved in the linking process, the within-grade variability patterns, and the examinee group differences, but they did not appear to be a function of the maintenance approach implemented.

The findings from this study provide important empirical guidance to practitioners on how the vertical scale can be maintained, once a vertical scale is established. If the

desired output of a maintained scale is to continue to capture the characteristics of the established scale in terms of grade separation, within-grade variability, and growth implications, the vertical approach appears to be marginally better in achieving these goals. However, the differences observed across three forms are marginal, and in some cases neither approach demonstrates superiority in preserving the same patterns at the baseline scale. Both approaches are able to reasonably well capture the trend of the baseline scale, at least across three forms.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER I
INTRODUCTION

Vertical scaling, sometimes called vertical linking, is a method of developing a common metric across grade levels that allows comparisons among scores from tests that differ in difficulty but that are intended to measure similar constructs. One advantage of placing achievement tests administered in consecutive grades on a vertical scale is that it permits educators to make inferences about student achievement or growth across grades.

Recent political and policy changes have also contributed to the importance of high quality test equating through accurate vertical scaling. In response to the release of the blueprint for the *Reauthorization of Elementary and Secondary Education* (U.S. Department of Education, 2010), American states and their school districts are increasingly implementing accountability policies that focus on not only levels of student achievement, but also on student growth. Even though the reauthorization of ESEA doesn't require states to construct vertical scales or report longitudinal growth, vertical scaling -- the process of linking different levels of tests measuring the same construct onto a common scale – can make it easier to evaluate growth from one grade to the next (Harris, Hendrickson, Tong, Shin, & Shyu, 2004).

Once a vertical scale for an assessment has been constructed, it also needs to be maintained over different forms and over time. If new forms of the assessment are developed in years subsequent to the baseline year, then an equating design must be

incorporated as part of the new form implementation. Issues such as data collection designs, equating methodologies, and examinee sample characteristics need to be considered while equating new forms to a vertical scale (Harris, 2007). The issue on how to maintain a vertical scale over time has been largely ignored in the literature until recent years. Almost all the research on this topic has focused on whether vertically linking one form of a multi-level form to another is preferable to maintaining a multi-level scale by horizontally equating individual levels across two forms or over two years. (cf. Cao, Li, & Hendrickson, 2007; Hoskens, Lewis, & Patz, 2003; Tong & Kolen, 2008; Tong & Kolen, 2009; Wang & Harris, 2009a; Wang & Harris, 2009b; Tomkowicz, Zhang, & Yen, 2010). These recent studies compared different approaches for maintaining a vertical scale over two forms and yielded similar results given the specified conditions and constraints of the studies. However, no consensus exists as to which scale maintenance approach is preferable in any given situation over two forms. In addition, it is possible that there may be more drift over a chain of multiple forms. Therefore, maintaining a vertical scale over multiple forms is the central theme of this dissertation, as drift may be small and not noticeable until a longer chain is examined. Assuming that an original vertical scale has been developed, this dissertation compares two scale maintenance approaches: *horizontal equating within grade* and *horizontally linking two vertical scales*. Issues such as within-grade variability patterns and examinee sample characteristics are taken into consideration in the application of the two scale maintenance approaches. This dissertation will shed light on differences in the resulting proficiency estimates by using

these two approaches in supporting scale stability across grade within form, within grade across forms, and across grades and across forms, and will investigate under which conditions one approach is preferable to the other.

Construction of Vertical Scales

Constructing a vertical scale is a complicated process. Decisions have to be made with respect to the definition of growth, scaling design, statistical methods, type of scales, and so forth (Briggs, Weeks, & Wiley, 2008; Chin, Kim, & Nering, 2006; Harris et al., 2004; Harris, 2007; Kolen, 2003; Lin & Doran, 2011; Liu, 2010; Proctor, 2008; Tomkowicz, Zhang, & Yen, 2010;Tong, 2005). Different decisions may lead to somewhat different vertical scales. Research shows that vertical scaling is design-dependent (Harris, 1991 & 2007; Kolen & Brennan, 2004; Tong & Kolen, 2007), group-dependent (Harris & Hoover, 1987; Lin & Doran, 2011; Powers, Turhan, & Binici, 2012; Skaggs & Lissitz, 1988; Slinde & Linn, 1979; Tong & Kolen, 2008), and method-dependent (Chin, Kim, & Nering, 2006; Harris, 2007; Kolen, 1981; Skaggs & Lissitz, 1986; Tong & Kolen, 2007;Topczewski, 2012). No single set of procedures has been proposed in the literature as being “best” in establishing a vertical scale.

Constructing vertical scales is the first step in investigating the process of maintaining an established vertical scale to support scale stability. It involves conceptual, technical, and implementation decisions (Harris et al., 2003; Harris, 2007; Kolen & Brennan, 2004; Proctor, 2008; Tong, 2005), such as the nature of growth, data collection design, and scaling methods.

Conceptual Issue: the Nature of Growth

The primary reason for creating vertical scales is to measure learning across time. Without an understanding of the nature of growth, it is not possible to clearly evaluate whether a vertical scale is functioning as it should. The nature of growth is primarily a developmental issue that can be informed by psychological research and theory and influenced by how the educational curriculum is implemented (Harris et al., 2004; Harris, 2007). It is the nature of growth and whether within-grade variance increases, decreases, or remains constant that are key issues in the debate over scale shrinkage in the vertical scaling literature (e.g., Becker & Forsyth, 1992; Camilli, 1988; Pommerich & Thissen, 1998; Yen & Burket, 1997). Under different scaling methods, the pattern of within-grade scale variability may vary over grade. Harris (2007, p.239) stated:

> *One problem in trying to address the issue of defining growth is that test publishers rarely make the information explicit. It seems that most definitions are determined operationally, based on a combination of empirical data, the test development process, and preconceptions regarding the nature of growth. For example, a practitioner who believes within-grade variance should remain constant over grades might not develop test specifications or a data collection design with this in mind, but might reject scaling methods that resulted in large changes in within-grade variance over grades.*

It is recommended that researchers and practitioners developing vertical scales should make as explicit as possible their assumptions regarding the nature of growth, and how these affect the developing of the scale (Harris, 2004). In this dissertation, the pattern of within-grade scale variability is considered as one of the conditions in developing a vertical scale.

Technical Issues

Constructing a vertical scale requires decisions about technical issues, such as data collection design and scaling methods.

Data Collection Designs

Different data collection designs can be used to create vertical scales. Two commonly used designs are the scaling test design and the common item design (Kolen & Brennan, 2004). In the scaling test design, a scaling test is constructed to be of a length that can be administered in a single setting and that spans the content across all of the grade levels. Students in multiple grades are administered the same scaling test. Under a common item design, each test level is administered to examinees at the appropriate grade. The link for test levels is established through students' performance on the items that are common between any pair of adjacent grades. Between these two designs, the common item design is easier to implement when the test contains items that are common to adjacent levels. In this case, the common-item design is implemented using standard administration conditions with the standard test (Kolen & Brennan, 2004). The scaling test design is more difficult, since it requires construction of a scaling test, and a special administration in which the scaling test is administered to students in each grade. In this dissertation, the common-item design is employed for constructing vertical scales. The common items between the two adjacent grades are denoted as vertical common items.

Scaling Methods

Once the nature of growth and data collection design are determined, the next consideration is the scaling method to be used to develop a vertical scale. IRT scaling is perhaps the most utilized method to scale achievement tests, both vertically and horizontally. Under an IRT model, items are assumed to have a set of parameters and the number of parameters for items depends on the IRT model chosen. The chosen model usually links the probability of an item with certain parameters being answered correctly to the examinee's proficiency on the construct(s) through a logistic function. IRT models also require that the items measure the same construct(s) as well as satisfy the assumption of local independence.

Implementation Issues

When implementing IRT for vertical scaling, scale results may be affected by several factors: dimensionality, model choices (the Rasch Model versus the Three-Parameter Logistic Model), choice of item parameter linking methods (mean-mean, mean-sigma, Stocking-Lord, etc.), and calibration methods (concurrent, separate, etc.).

Model Choices

A number of different IRT models could be used for scaling: unidimensional and multidimensional IRT models, dichotomous and polytomous IRT models, parametric and nonparametric IRT models, and so forth. Due to the estimation complexities, the multidimensional models have not been used widely in the practice of vertical scaling. Thus, only unidimensional dichotomous IRT models are considered in this dissertation.

For dichotomously scored items, the most common models include the One-Parameter Logistic Model (the Rasch Model) and the Three-Parameter Logistic Model (3PLM, Lord, 1980). Other models exist for polytomously-scored items or for items based on a common stimulus, such as the Graded Response Model (Samejima, 1997), Bock's Nominal Model (Bock, 1997), and the Generalized Partial Credit Model (Muraki, 1997). Mixed findings exist in the literature on choosing IRT models for constructing a vertical scale. In different studies, the Rasch Model was found to be both acceptable (e.g., Schulz, Perlman, Rice, & Wright, 1992) and unacceptable (e.g., Phillips, 1983). Some studies have also demonstrated that the 3PLM produced better results than the Rasch Model (e.g., Kolen, 1981; Lord, 1977; Marco, 1977; Marco, Peterson, & Stewart, 1983; Loyd & Plake, 1987; Patience, 1981; Skaggs & Lissitz, 1986). Therefore, the 3PLM is used to construct vertical scales in this dissertation.

Concurrent and Separate Calibrations

Several methods for establishing a common scale across grades within a content area using the common-item design have been described (Hanson & Beguin, 2002; Kim & Cohen, 1998: Patz & Hanson, 2002; Larkee, Lewis, Hosken, Yao, & Haug, 2003). Two of these methods are widely used in vertical scaling: *separate calibration* and *concurrent calibration*.

Under the common-item design, separate calibration is accomplished in two steps. First, item parameters and examinee ability are estimated for each grade separately, resulting in each grade being on its own separate scale. The scale for one of the grades is

identified as the base scale. In the second step, a linear transformation can be utilized to link the two sets of item parameter estimates for the base grade and the adjacent grade and place them on the same scale. The transformation constants can be estimated from vertical common items. There are four methods that can be used for scale transformation: mean/sigma (Macro, 1977), mean/mean (Loyd & Hoover, 1980), Haebara (Haebara, 1980), and Stocking-Lord (Stocking & Lord, 1983). To the author's knowledge, the first three methods are not currently used operationally and therefore will not be utilized in the current study. Thus, the Stocking-Lord approach is used in this study to obtain the transformation constants. This second step is repeated for each adjacent grade until all grades are on the common base scale through a linking chain.

For concurrent calibration, item parameters and examinee ability are estimated for all the grades combined, resulting in just one scale and requiring only one computer run. Concurrent calibration is more efficient and uses more information than separate calibration. However, with all grades combined, the assumption of unidimensionality may be violated (Kolen, 2006; Kolen & Brennan, 2004).

As mentioned above, many choices need to be made when IRT methodology is used, adding more complexity to constructing vertical scales. In this dissertation, separate calibration and different patterns of within-grade variability are used in developing sets of vertical scales.

Scale Maintenance

Rationale

Recent scholarship has frequently addressed the issue of how to maintain a vertical scale over time, mainly focusing on whether vertically linking one form of a multi-level form to another is preferable to maintaining a multi-level scale by horizontally equating individual levels (cf., Cao, Li, & Hendrickson, 2007; Hoskens, Lewis, & Patz, 2003; Tong & Kolen, 2008; Tong & Kolen, 2009; Wang & Harris, 2009a; Wang & Harris, 2009b). Hoskens, Lewis, and Patz (2003) noted that a year-to-year or form-to-form vertical scale equating design should foster comparability of the scale in three ways.

First, the process of maintaining an established vertical scale should foster comparability across forms within grade. The measure of growth in this way does not require a vertical scale. Performance comparisons within each grade can be conducted between the two forms. Second, a form-to-form vertical scale equating design should support comparability across grades within form. Scale stability across grades within year is vital for meaningful comparisons between mean scores across grades. Third, a form-to-form vertical scale equating design should also allow comparability across grades and forms. It is this comparability that allows individual student growth to be measured across forms by tracking the change in the students' scale scores from one administration to another.

If an original vertical scale already exists, simple horizontal equating is always an option, but caution needs to be taken to prevent scale drift over time (Hoskens et al., 2003; Tong & Kolen, 2008). It is a complicated process to design a form-to-form vertical scale equating methodology for scale maintenance. When an IRT model is involved, issues such as equating methodologies, and examinee sample characteristics need to be considered in equating new forms to a vertical scale (Harris, 2007), in combination with various choices of within-grade variability, IRT model, and calibration method in constructing the vertical scale. Recent studies compared different approaches for maintaining a vertical scale over two forms and yielded similar results for the conditions specified in the studies (cf., Cao et al., 2007; Hoskens et al., 2003; Tong & Kolen, 2008; Tong & Kolen, 2009; Wang & Harris, 2009a; Wang & Harris, 2009b). As this scholarship demonstrates, no consensus exists as to which scale maintenance approach is preferable in any given situation over two forms. In addition, more drift over a chain of multiple forms is possible. Therefore, the major purpose of this dissertation is to compare how different the resulting proficiency estimates are by using two scale maintenance approaches, in combination with different within-grade variability patterns and examinee sample characteristics, given that a vertical scale is established on the baseline form.

Data Collection Design

Linking across three forms in vertical scaling involves an equating chain through which new forms are equated to an established vertical scale. In this dissertation, the three forms are denoted as Form Y (base form), Form Z (interim form), and Form X (new

form). Each form consists of three level tests (i.e., one for each of Grades 10-12). The scales on Form X are linked to the established vertical scale on Form Y through the interim Form Z (X to Z to Y). Under each of the various combinations for constructing vertical scales, three sets of vertical scales are established spanning from Grade 10 to Grade 12: one for the baseline Form Y, one for the interim Form Z, and one for the new Form X. In this study, linking across three forms in vertical scaling under the common-item nonequivalent groups design is considered. There are horizontal common items for the same grade level across Form Y and Form Z, and Form Z and Form X, and vertical common items between adjacent grades within the same form. For example, for the Grade 10 level test across forms, Form Z shares horizontal common items both with Form Y and with Form X. However, Form Y and Form X do not share common items. Similarly, there are vertical common items between Grade 10 and Grade 11 for a given form. The horizontal common items for the same grade between forms are used to horizontally equate the tests via an equating chain: Form X to Form Z to Form Y. The vertical common items between two adjacent grades for a given form are used to establish a vertical scale spanning all grade levels. The IRT 3PLM is used in the process of scale maintenance.

Calibration Method

Under the common item nonequivalent groups design the three forms of a test with common items are administered to three different samples. The group of examinees taking the baseline Form Y is denoted as Group 1; the group of examinees taking the

interim Form Z is denoted as Group 2; and the group of examinees taking the new Form X is denoted as Group 3. Each group contains Grades 10-12. To link Form X to the baseline vertical scale on Form Y via Form Z, separate estimation is used to place proficiency and item parameter estimates on a common scale through horizontal common items. The separate calibration procedures in equating Form X to the established vertical scale on Form Y across the three forms are similar to those used in vertical scale construction (see details in the above section). The Stocking-Lord approach (Stocking & Lord, 1983) is used to estimate transformation constants for scale maintenance. The only difference is that horizontal common items are used for equating across forms while vertical common items are used for linking adjacent grades.

Scale Maintenance Approaches

Given that the baseline vertical scale is established on Form Y, there are two possible ways to maintain a vertical scale across the three forms: (1) construct the original vertical scale on Form Y and maintain it through horizontal equating within grade across forms or (2) construct separate vertical scales for the three forms and horizontally link the two vertical scales of Form Y and Form X through the interim Form Z.

Horizontal Approach

One advantage of the horizontal approach is that while vertical common items are needed in the baseline Form Y to establish the vertical scale no vertical common items are needed in Form Z and Form X. Horizontal common items for the same grade between

forms are used to establish the linking relationship. Through an equating chain, the parameter estimates for each grade level on Form Z are placed onto the established vertical scale on Form Y. Four steps are involved under this horizontal approach. First, a baseline vertical scale is established on Form Y using vertical common items between adjacent grades. Second, for a given grade, the transformation constants are obtained using the Stocking-Lord method with horizontal common items between Form Z and Form Y. Third, for each grade, the transformation constants are obtained using the Stocking-Lord method with horizontal common items between Form Z and Form X. Finally, the results of the level test on Form X are transformed to the established vertical scale on Form Y. A total of two transformations are involved.

Vertical Approach

Under this approach, three separate vertical scales are developed using vertical common items for each of the three forms. Next, the vertical scale developed on Form X is then linked back to the vertical scale on Form Y via the interim Form Z through an equating chain. This approach is more complicated than the horizontal approach.

To achieve the linking of the two vertical scales on Form X and Form Y via Form Z, horizontal common items are used to identify the linking relationship. They are the same sets of horizontal common items as those used in the horizontal approach; under the vertical approach, these items are used to link vertical scales instead of linking tests within the same grade. To obtain the corresponding linking transformation constants between the vertical scales on Forms Y and Z and between the vertical scales on Forms Z

and X, the parameter estimates for each grade level on all the three forms need to be placed onto their respective vertical scales. The transformation constants from the previous horizontal approach cannot be used directly, because those constants are on the same scale as each of the grades but not on the vertical scale. Two sets of transformation constants are obtained using the Stocking-Lord method based on all the horizontal common items through Grades 10-12, between Forms Y and Z and between Forms Z and X. Finally, the vertical scale on Form X is linked to the established vertical scale on Form Y through an equating chain, using the two sets of transformation constants. A total of two transformations are involved.

Examinee Sample Characteristics

In equating, two groups of respondents are generally involved, one responding to the old form and one responding to the new form to be equated to the old form. Harris (1993) reviewed previous research to investigate the impact of sample characteristics on equating results (cf. Harris, 1993) and found a general consensus that the more alike the samples are the better. However it was unclear how different is too different or how much worse an equating is when the examinees taking Form X are not considered to be equivalent to those taking Form Y.

While examinee sample characteristics have been studied quite a bit in the equating literature (e.g., Eignor & Schmitt, 1989; Kolen & Brennan, 2004; Lin & Doran 2011; Power, Turhan, & Binici, 2012; Skaggs, 1990a & 1990b; Stocking & Eignor, 1986), it has been virtually ignored in the vertical scale maintenance scholarship.

Therefore, in this dissertation *examinee sample characteristics* are considered as one of the factors assumed to have potential impact on the resulting scales using different combinations of scale maintenance approaches.

Evaluation Criteria

A simulation study is conducted for this dissertation. A strong advantage to conducting a simulation study is that objective criteria exist to evaluate the results since true item parameters and student proficiencies are known. In this dissertation, two vertical scale maintenance approaches are introduced and compared to examine their practical impact on the resulting proficiency estimates, in combination with within-grade variability patterns and examinee sample characteristics. To investigate the characteristics of these scale maintenance methods, evaluation criteria need to be identified.

Three criteria are used to evaluate the overall fit of the different combinations of scale maintenance approaches, within-grade variability patterns, and examinee sample characteristics, given an established vertical scale. The first is the correlation between the true proficiencies and the estimated proficiencies of the examinees. The correlation used is the Pearson (1896) product-moment coefficient. To investigate the adequacy of the two scale maintenance approaches under various conditions of within-grade variability patterns and examinee group differences in recovering the true examinee proficiencies, the Root Mean Square Error (RMSE) and the Mean Absolute Difference (MAD) are used as the second and third criteria, showing the extent to which the estimated proficiency values match the true values using different combinations to link across forms in vertical scaling.

To capture the growth trends established by the combinations of the scale maintenance approaches, three properties proposed by Kolen and Brennan (2004) were employed: grade-to-grade growth, grade-to-grade variability, and separation of grade distributions. For the grade-to-grade growth, the mean proficiency difference between adjacent grades was calculated to check the overall growth of the students from lower grade to higher grade. For the grade-to-grade variability, the differences in the standard deviations of the resulting proficiency estimates between the adjacent grades were calculated. The third property, the separation of grade distributions, is examined using an index proposed by Yen (1986) for the effect size of grade-to-grade differences.

To compare the performance of the two scale maintenance approaches under different conditions of within-grade variability pattern and examinee sample characteristics, the ratios of the RMSEs and the ratios of MADs between the vertical approach and the horizontal approach within form are used, as well as the ratios of the RMSEs and the ratios of MADs between the new Form X and the interim Form Z via each approach.

Research Questions

Maintaining vertical scales across forms is an involved process with many decisions. Given that a vertical scale for the original form is established, a form-to-form vertical scale equating design should foster three types of comparability -- within grade across forms, across grades within form, and across grade and across forms. Recent studies mentioned earlier compared different approaches of maintaining a vertical scale over two forms and yielded similar results in the conditions examined in the literature

(cf. Cao, Li, & Hendrickson, 2007; Hoskens, Lewis, & Patz, 2003; Tong & Kolen, 2008; Tong & Kolen, 2009; Wang & Harris, 2009a; Wang & Harris, 2009b; Tomkowicz, Zhang, & Yen, 2010). No consensus exists as to which scale maintenance approach is preferable in any given situation over two forms. In addition, it is possible that there may be more drift over a chain of multiple forms. Therefore, linking over three forms in vertical scaling is the central theme of this dissertation, as drift may be small and not noticeable until a longer chain is examined. Given an original vertical scale is developed, two scale maintenance approaches -- *horizontal equating within grade* and *horizontally linking two vertical scales* -- are introduced and compared in this dissertation. Two factors –within-grade variability patterns (constant, decreasing, and increasing) and examinee sample characteristics (small and large sample differences) -- are taken into consideration in the application of the two scale maintenance approaches. This dissertation is intended to compare how different the resulting proficiency estimates are by using these two approaches in supporting scale stability across grade within form, within grade across forms, and across grades and across forms, and to investigate under which conditions one approach is better than the other. Investigating the similarity and dissimilarity of the resulting scales via these two approaches is the focus of this dissertation.

In the process of scale maintenance, a vertical scale is first established for the baseline form (Form Y). In developing a vertical scale for the baseline form, the factor of within-grade variability patterns is considered: constant across grade, decreasing across grade, and increasing across grades. Thus, three sets of vertical scales are constructed for

the baseline form. In addition, in generating data through Grade 10 to Grade 12 across three forms, the factor of examinee sample characteristics is considered: small difference and large difference. Thus, in combination of three within-grade variability patterns and two types of examinee sample characteristics, six conditions of proficiency distributions are used in simulating the data and the performance of the two scale maintenance approaches.

In sum, this study uses existing item data to build different combinations of simulated tests and evaluates how different the two scale maintenance approaches perform on providing equivalent resulting proficiency estimates on the new form through a linking chain of three forms. The manipulated simulations have the following facets:

A. Three forms: Form Y (the original baseline form), Form Z (the interim form for linking), and Form X (the new form to be linked to Form Y via From Z);
B. Three grade levels: Grade 10, Grade 11, and Grade 12, nested in the three forms (A above).
C. Two levels of examinee group differences, small and large (d_{mean}= 0.25 and d_{mean}= 0.50), nested in grade levels (B) nested in forms (A)
D. Three patterns of within-grade variability (constant, increasing, decreasing), nested in group level differences (C) nested in grade levels (B), nested in forms (A).

These procedures result in six conditions (C by D) to be considered in the simulation. For each of the six conditions, a set of nine data combinations is generated

simultaneously, which consists of three level tests for each of the three forms. Thus six different sets of data combinations are generated for the six conditions to be investigated.

The research questions to be investigated for this dissertation are:

1. How adequate are the two scale maintenance approaches in recovering the true examinee proficiencies, under the six conditions of proficiency distributions, considering the factors of within-grade variability patterns and examinee sample characteristics?
2. What are the effects of the two scale maintenance approaches on the resulting proficiency estimates and growth interpretations, under the six conditions of proficiency distributions, considering the factors of within-grade variability patterns and examinee sample characteristics?
3. Under which conditions of proficiency distributions does one scale maintenance approach provide more equivalent resulting proficiency estimates than the other, considering the factors of within-grade variability patterns and examinee sample characteristics?

CHAPTER II
LITERATURE REVIEW

As discussed in Chapter I, maintaining vertical scales across forms is a complicated process, which involves both constructing a vertical scale and equating new forms to an existing vertical scale. In the first section of this chapter, literature review will focus on scale maintenance approaches that have been discussed only recently by the psychometric literature. In the second section, factors that are likely to affect the results in equating new forms to a vertical scale are reviewed, such as model choices, calibration methods, equating methods and examinee sample characteristics. Although these factors have been generally explored in the literature on horizontal equating, it is not clear how these factors affect the results in equating new forms to a vertical scale. In addition to equating new forms to a vertical scale, it is necessary to create an existing vertical scale as the base scale. The existing literature suggests that constructing a vertical scale is a very complex process that is affected by many factors. Therefore, in the last section, factors such as patterns of within-grade variability across grades, model choices, and calibration methods will be explored as part of the literature review.

Scale Maintenance

Rationale

The issue on how to maintain a vertical scale over time has been addressed recently, mainly focusing on whether vertically linking one form of a multi-level form to

another is preferable to maintaining a multi-level scale by horizontally equating individual levels (cf. Cao et al., 2007; Hoskens et al., 2003; Tong & Kolen, 2008; Tong & Kolen, 2009; Wang & Harris, 2009a; Wang & Harris, 2009b; Tomkowicz, Zhang, & Yen, 2010).

One paper that was very useful in developing a list of issues on equating new forms to a vertical scale was one by Harris (2007) entitled “Practical issues in vertical scaling”. In this paper, Harris raised several questions as examples of what she thought needed to be addressed in scale maintenance, one of which influenced the current dissertation: should new Grade 3 forms be equated to the original Grade 3 form or should there be an attempt to link the entire range of Grade K to Grade 8 forms to the original set of forms on which the scale was set.

Hendrickson, Tong, Shin, and Shyu (2004) claimed that research has tended to focus on the initial development of vertical scales, not on their maintenance over time. They recommended that procedures which lead to a more stable scale over time should be identified and criteria for evaluating the long term use of scales need to be developed. In a study of comparisons of methodologies and results in vertical scaling, Tong (2005) also mentioned the necessity of conducting further research on the issue of maintaining vertical scales across years.

Hoskens, Lewis, and Patz (2003) noted that year-to-year or form-to-form vertical scale equating designs should foster comparability of the scale in three ways. First, the process of maintaining an established vertical scale should foster comparability across

forms within grade. The measure of growth in this way does not require a vertical scale, and performance comparisons within each grade can be conducted between the two forms. Second, form-to-form vertical scale equating designs should support comparability across grades within form. Scale stability across grades within year is vital for meaningful comparisons between mean scores across grades. Third, form-to-form vertical scale equating should also allow comparability across grades and forms. It is this comparability that allows individual student growth to be measured across forms by tracking the changes in the students' scale scores from one administration to another.

Given that an original vertical scale exists, simple horizontal equating is always an option, but caution needs to be taken to prevent scale drift over time (Hoskens et al., 2003; Tong & Kolen, 2008). Via simple vertical scaling, the comparability across grades within form may be achieved while the other two types of comparability cannot be supported (Hoskens et al., 2003). It is a complicated process to develop a form-to-form vertical scale equating scenario for scale maintenance.

Scale Maintenance Approaches

As mentioned above, the issue of scale maintenance methods has been largely ignored in the literature until recent years. This section reviews the scale maintenance methods.

Hoskens, Lewis, and Patz's Study (2003)

Hoskens, Lewis, and Patz (2003) considered the issue of maintaining a vertical scale across years using a baseline vertical scale. The tests spanned seven levels,

corresponding to Grade 4 through Grade 8. Each level consisted of items in common with 1) previously administered forms of the assessment for which parameters on the base vertical scale existed (horizontal anchors) and 2) adjacent levels of the new form of the assessment (vertical anchors) for which parameter estimates did not yet exist. In their study they used the 3PLM for the analysis of multiple-choice items, and the Two-Parameter Partial Credit Model for the analysis of constructed-response items. The Stocking-Lord procedure was used for equating and linking. Four different methods were examined for maintaining vertical scales:

1. Horizontal equating within each grade. Given that a baseline vertical scale was established in the assessment of the previous administration, a standard approach to maintain vertical scales was to horizontally equate the assessment of the current administration to the assessment of the previous administration for each grade level. Under this approach, the horizontal anchor items within a given grade across the two assessments were used in the equating procedure to obtain the corresponding transformation coefficients. These transformation coefficients within a given grade were used to place the parameter estimates for the current assessment onto the established vertical scale for the previous assessment. The horizontal equating procedure was carried out separately within each grade level.
2. Augmented method that used both vertical and horizontal anchors. The augmented anchor sets procedure in a sense combines horizontal equating and vertical linking by using both types of anchor items simultaneously to achieve

the link to the vertical scale that was established previously. Anchor items used in this approach were both horizontal and vertical anchors, i.e., items that were common with the on-level assessment of the previous administration and items that were common with the assessments of adjacent higher and lower levels. Since the vertical anchor items (most likely) might have been administered in the current administration only, the target scale item parameter estimated for the two types of anchor items initially were in different metrics. Therefore, an initial equating was done to put the vertical anchors on the same metric as the horizontal anchors. After the initial (horizontal) equating for each level, more items were added to the pool of calibrated items that could be used as anchors in the augmented anchor sets. Then a second step that involved linking with the augmented anchor set was carried out for each level.

3. Concurrently develop another vertical scale with the new forms. The concurrent horizontal equating approach required that the assessment in the current administration was vertically scale first, followed by a horizontal equating of this newly developed vertical scale to the previously existing vertical scale. In this approach, a vertical scale was established anew for the new assessment, using the concurrent calibration approach. This vertical scale for the new forms was then linked back to the previously established vertical scale using concurrent horizontal equating, using all horizontal anchors over all grades simultaneously.

4. Separately develop another vertical scale and then link the two vertical scales. In this approach a vertical scale was established first for the new assessment, using separate calibration (and chained linking using the vertical anchors, Patz & Hanson, 2002; Karkee, Lewis, Yao & Haug, 2003). This vertical scale for the new forms was then linked back to the previously established vertical scale using concurrent horizontal equating, using all horizontal anchors over all grades simultaneously.

Cao, Li, and Hendrickson's Study (2007)

In the study conducted by Cao, Li, and Hendrickson (2007), three procedures for scale maintenance were investigated using simulated data across Grades 3 to 6:

1. Separate calibration + separate horizontal equating. Item and person parameters for each grade-level test of the new form were estimated separately. Separate horizontal equating of the new form to the old form scale was done using the Stocking-Lord method with the horizontal common-item sets for each grade-level test. With the scale of each grade-level test on the new form converted to that of the corresponding old form grade-level test, the scale on the new form was linked to the established vertical scale on the old form.
2. Separate calibration + horizontal equating + vertical scaling. In this procedure, item and person parameters for each grade-level test of the new form were estimated separately. Horizontal equating was conducted to place the Grade 3 new form test onto the scale of the Grade 3 old form test with the Stocking-Lord

method, using the horizontal common items. The new form vertical common-item sets were then used to place item and person parameters from the higher grades onto the Grade 3 scale using the Stocking-Lord method. Thus, all the new form level tests were converted to the same scale as the old form level tests.

3. Concurrent calibration + horizontal equating. In this procedure, item and person parameters for all grade-level tests of the new form were estimated concurrently. Horizontal equating was conducted to place the Grade 3 new form test onto the scale of the Grade 3 old form level test with the Stocking-Lord method using the horizontal common items. The item and person parameter estimates of the new form were then converted to the old form vertical scale.

In addition to investigating these three procedures, the study also examined sample size as a possible factor affecting results. The two sample sizes used for the study were 500 for each grade of each form for a total of 4000 simulees and 1000 for each grade of each form for a total of 8000 simulees. However, the results of the study indicated that the two different sample sizes did not affect the performance of the three scale maintenance procedures.

Tong and Kolen's Studies (2008 & 2009)

Tong and Kolen (2008) analyzed the data collected from a large-scale state assessment program over a period of two years with embedded vertical linking items and equating linking items. They fit the Rasch Model and the Partial Credit Model to the data for analysis and scale explored two maintenance approaches, *horizontal equating* and

equating two vertical scales, to observe their impact on the resulting scale and growth interpretation.

1. Horizontal equating. Under this approach, vertical linking items were only needed in the first year to establish the vertical scale. Common items for the same grade between two years were used to establish the equating relationship. Through equating, the scores from future administrations were placed onto the established vertical scale. Item parameter estimates for the horizontal common items in the two years were obtained through separate calibration. The average of the item parameters estimated for these common items for both years was computed and the difference was obtained to serve as the equating constant to link the two administrations. To place the second year results onto the first year scale, the corresponding equating constant was added to parameter estimates for each of the grade levels. No further adjustment was needed because the first year was already on the baseline vertical scale. Through horizontal equating, the second year results were also placed onto the same scale.
2. Equating two vertical scales. This approach established multiple vertical scales and then equated the vertical scales. Under this approach, two separate vertical scales (one for each year) were constructed. After construction of the two separate vertical scales, the vertical scale developed for the second year was placed onto the vertical scale for the first year. To achieve this linking of the two vertical scales, horizontal equating was applied. The equating constants for each

grade level across the two years were obtained with the mean/mean method using the horizontal common items within a given grade, resulting in six sets of equating constants that were computed (for the corresponding grade levels). The average of all the equating constants was used to link the two vertical scales.

Tong and Kolen (2009) conducted a study to have a further look into the maintenance of vertical scales. They used the same data as their previous study and the same two scale maintenance approaches. Under the 3PLM and the generalized partial credit model, they looked at two subject areas (Math & ELA) and found similar results for horizontal equating and vertical linking.

Wang and Harris' Studies (2009a & 2009b)

Wang and Harris (2009a & 2009b) used simulated data and the 3PLM approach to have a known criterion when they looked at maintaining a vertical scale over two forms using a common-item nonequivalent groups design. They investigated two methods of maintenance, *horizontal equating within grade* and *horizontally linking two vertical scales*, by considering the patterns of within-grade proficiency variability in the process of constructing a baseline vertical scale. The findings indicated that the scale maintenance approaches were, to some extent, sensitive to the way the baseline vertical scale was established regarding different patterns of within-grade proficiency variability across grade. They found that the two methods were adequate in recovering the true examinee abilities. They indicated that it was hard to make a recommendation based on the mixed results observed in their studies.

Tomkowicz, Zhang, and Yen's Study (2010)

Tomkowicz, Zhang, and Yen (2010) investigated four methods using the 3PLM: concurrent calibration + horizontal on-grade level equating, concurrent calibration + vertical and horizontal concurrent equating, separate calibration + horizontal on-grade level equating, and separate calibration + vertical and horizontal concurrent equating. They used data spanning three years for a large scale state assessment program in Grades 3-8 English Language Arts (ELA) and Mathematics Tests. On-grade level ELA and Mathematics operational test items were administered to all Grades 3-8 students during regular administrations in each year of the study. The off-grade level linking items were selected from the adjacent grade operational assessments and were administered to samples of the same students during a separate administration approximately two weeks after the operational test administration. The on-grade level items were used as anchor items in horizontal linking scales across years, and the off-grade level items were used as anchor items in constructing vertical scales across Grades 3-8 for each year. They found that growth patterns across grades and across years appeared to be a function of the maintenance method implemented but appeared not to be a function of the baseline scale being maintained. Their findings indicated that the vertical/horizontal equating method produced more grade-to-grade growth and larger within-grade variability than the horizontal on-grade level equating method, while the latter seems to preserve the characteristics of the baseline scales better.

Comparisons of the Scale Maintenance Methods

These research studies on the comparison of the scale maintenance methods yielded mixed results. Hoskens, Lowis and Patz (2003) concluded that using different methods to maintain a vertical scale in years subsequent to the baseline year would affect the results in non-trivial ways. On the other hand, Tong and Kolen's (2008 & 2009) found similar results for the two methods they used. The three different procedures in Cao et al.'s (2007) study also showed good performance in capturing the true growth trend of the simulated data. Wang and Harris' (2009a & 2009b) studies also indicated the adequacy of the two methods they used in recovering the true examinee abilities. Tomkowicz, Zhang, and Yen (2010) also concluded that different maintenance methods used in their study did not produce dramatically different scales, regardless of whether the baseline vertical scale was created using concurrent calibration or chain linking.

In the Hoskens et al. study (2003), three methods indicated growth between grades in the scale scores, but the horizontal method showed essentially no growth in mean scale scores from Grade 5 to Grade 6. In terms of grade-to-grade variability, the variability was relatively flat (augmented method) or moderately decreasing over grades (horizontal method). The other three methods, for which the vertical scale was re-established across grades prior to horizontally anchoring to the scale of baseline year, indicated significant increases in variability across grades, even though the raw score distributions did not reflect such patterns.

In Tong and Kolen's (2008 & 2009) studies, the results of linking vertical scales were very similar for the two methods. They concluded that with the dataset used, it would not have mattered much which maintenance model was used to maintain the vertical scale for the second year. On one hand, it was suggested that the horizontal equating approach is the more straightforward and is easier to apply in practice, while equating two vertical scales is more complicated, demanding vertically linked items be administered in multiple years. Given all these results, it might still be preferable to consider multiple vertical scales in situations where linking is done over multiple years.

Cao et al. (2007) noted that the differences among the three procedures of scale maintenance were so small that one could not claim that any one of the procedures is superior to the other two. Their findings indicated that the two different sample sizes (500 and 1000) did not affect the performance of the three procedures. They concluded that any of the three procedures could successfully maintain the vertical scale if the new form was well designed so that it shared similar constructs and statistical characteristics with the old form, and the common items were representative and sufficiently numerous. They further suggested that additional investigations explore the impact of proficiency variance, number of common items, different IRT models, and calibration methods on the procedures of scale maintenance.

Wang and Harris (2009a & 2009b) indicated that the scale maintenance approaches are, to some extent, sensitive to the way the baseline vertical scale is established regarding different patterns of within-grade proficiency variability across

grades. They suggested further research comparing the scale maintenance methods. Issues related to model choices, calibration methods, equating methods, and sample characteristics should also be taken into consideration in the application of scale maintenance approaches.

If the objective is to preserve the characteristics of the baseline year in terms of grade-to-grade growth and within-grade variability, Tomkowicz, Zhang, and Yen (2010) suggested that, the horizontal on-grade level equating method seems to be better than the vertical/horizontal equating method. However, they pointed out that the difference in preservation of these characteristics between methods is small.

There are similarities among these studies. In all three, IRT-based models were fit to the data for analyses and the common-item nonequivalent groups design was used. The dataset for each study consisted of both horizontal common items and vertical common items. Both *horizontal equating* and *equating two forms to an established vertical scale* were taken into consideration.

There are also dissimilarities among these three studies. First, different IRT models were used: the 3PLM and Two-Parameter Partial Credit Model in the Hoskens et al. study (2003); the 3PLM in Cao et al. (2007), in Wang and Harris (2009a, 2009b), and in Tomkowicz, Zhang, and Yen's (2010); the Rasch Model in Tong and Kolen (2008); and the 3PLM and the Generalized Partial Credit Model in Tong and Kolen (2009). Second, concurrent and separate calibration methods were both used for item parameter estimation in the Hoskens et al., Cao et al., and Tomkowicz et al. studies; however, only

the separate calibration method was considered in the Tong and Kolen and in Wang and Harris studies. Third, to obtain equating constants and transformation coefficients, the mean/mean approach was used in Tong and Kolen (2008), but the Stocking-Lord approach was used in all the other studies.

Given the various issues that were considered in the process of developing scale maintenance methods in the existing literature, the impact of different approaches to equating new forms to a vertical scale needs to be further investigated. The findings of the studies indicated that the scale maintenance approaches seemed to yield similar results over two forms in the conditions examined. However, more drift over a chain of multiple forms is possible. Changes over more than two forms need to be investigated, as drift may be small and not noticeable until a longer chain is examined.

This study will investigate linking over three forms in vertical scaling to see if more drift occurs. It will compare how different the resulting proficiency estimates are by using the different scale maintenance approaches and thus to investigate under which conditions of model choices and sample characteristics one approach is preferable to the other. Since IRT models and the common-item nonequivalent groups designs are the most used in practice, these two will be used in constructing vertical scales and in equating new forms to a vertical scale. The two scale maintenance methods to be considered are the horizontal approach (*horizontal equating within each grade*) and the vertical approach (*horizontally linking two vertical scales*). Details about these two methods are described in Chapter III.

Equating Scenarios

Since maintaining vertical scales involves equating new forms to a vertical scale, issues such as data collection designs, equating methodologies, and examinee sample characteristics need attention (Harris, 2007). The entire scope of equating issues is beyond the scope of the current study, which will focus on IRT models and the common-item nonequivalent groups design used in constructing vertical scales and in equating new forms to a vertical scale. The next section will focus on three important issues related to this study: the choice of IRT models, calibration methods, and examinee sample characteristics.

Choice of IRT Models

The literature describes a variety of IRT models: unidimensional and multidimensional IRT models, dichotomous and polytomous IRT models, parametric and nonparametric IRT models, and so forth. For dichotomously scored items, the most common models include the Rasch Model and the Three-Parameter Logistic Model (3PLM, Lord, 1980). Other models exist for polytomously-scored items or for items based on a common stimulus, such as the Graded Response Model (Samejima, 1997), Bock's Nominal Model (Bock, 1997), or the Generalized Partial Credit Model (Muraki, 1997). Due to estimation complexities, the multidimensional models have not been widely used in practice for horizontal equating and vertical scaling. Thus only unidimensional dichotomous IRT models are considered in this dissertation.

The Rasch Model and the 3PLM are widely used in equating context. These two models have been extensively studied in horizontal equating and vertical scaling literature, especially in the 1980s. This section reviews the research on the applications of these two IRT models in the horizontal equating context.

The Rasch Model

The Rasch Model (Lord, 1980) uses a logistic function to define the probability of an examinee with a given proficiency correctly answering an item. If θ_j stands for the ability parameter for an examinee *j* and b_i stand for the difficulty parameter for item *i*, then the probability of this examinee *j* answering item *i* correctly can be expressed by the following equation:

$$P_i(\theta_j) = \frac{1}{1 + \exp[-(\theta_j - b_i)]} \quad (1)$$

Yen and Fitzpartrick (2006) noted that Equation 1 is not only the simplest IRT model in use, it may be the most commonly used model. Besides the assumptions of unidimensionality and local independence, the Rasch Model also assumes that all the items have the same discrimination power and there is no guessing involved. It allows only item difficulty parameter to describe conditional probabilities of item responses for the entire range of abilities.

The findings of the Rasch Model application in the horizontal equating context are mixed. With tests of similar difficulty and samples of comparable ability, Rasch equating seemed to produce reasonable results (Anderson, Kearney, and Everett, 1968;

Rentz and Bashaw, 1977; Wright, 1968; Tinsley and Dawis, 1975; Whitely and Dawis, 1974). Skaggs and Lissitz (1986) noted that Rasch equating provided a major improvement over traditional methods, that is, item subsets could be tailored to specific groups. Whitely and Dawis (1974) suggested that ability estimates were not quite as invariant when tests were deliberately different in difficulty. Tinsley and Dawis (1975) pointed to potential problems that small sizes and samples widely different from one another in ability may cause when the Rasch model is used.

The 3PLM

The 3PLM (Lord, 1980) uses three parameters to characterize each item: the discrimination parameter (*a*), difficulty parameter (*b*), and the guessing or pseudo-guessing parameter (*c*). A logistic function is also used to define the probability of an examinee with a given proficiency correctly answering an item:

$$P_i(\theta_j) = c_i + \frac{1 - c_i}{1 + \exp[-Da_i(\theta_j - b_i)]} \quad (2)$$

where θ_j stand for the ability parameter for an examinee *j*; a_i, b_i, and c_i stand for the discrimination parameter, difficulty parameter and guessing parameter for item *i*; D is equal to 1.7.

Unlike the Rasch Model, the 3PLM accommodates guessing through *c* parameters and differing discriminating power of items through *a* parameters. A large value of an *a* parameter suggests high discriminating power of the item. The guessing parameter *c* is usually a value between 0 and .25 for multiple-choice type of questions (Lord, 1980).

Marco, Peterson and Stewart (1979) examined the Rasch model and the 3PLM under a variety of conditions, including random and dissimilar samples, internal and external anchor tests, and different types of criterion scores. They drew the following conclusions: when the anchor test was external and equal in difficulty across the two total tests, both the Rasch and the Three-Parameter Logistic Models performed well. When an external anchor test was used, the Rasch Model results were slightly better than those of the 3PLM.

Kolen (1981) expanded on the Marco et al. (1979) study by examining the two models. Kolen equated a new edition to a previous edition of the Iowa Tests of Educational Development (ITED) for two subtests, Vocabulary and Quantitative Thinking, using estimated true-score equating (Lord, 1980, p.199) and estimated observed score equating (Lord, 1980, p. 202). In addition, the Rasch model was modified so that the common discrimination (slope) was allowed to vary between the two tests being equated. The findings suggested a complex interaction between item content, difficulty level, and the models.

Skaggs and Lissitz (1988) compared the Rasch Model and the 3PLM equating methods for data generated from a unidimensional the three-parameter logistic model. Differences in mean test difficulty, mean test discrimination and degree of guessing were manipulated. The assumptions of the Rasch Model were systematically violated. Their results showed that the Rasch Model equating worked well when the same degree of

guessing was present in both tests. When test discriminations or the degree of chance scoring were unequal between the two tests, the Rasch model equating was inadequate.

Skaggs and Lissitz (1986) conducted a literature review on the IRT-model based methods in horizontal equating and concluded that neither model was universally superior to the other. Proponents of the 3PLM argued that the guessing parameter was needed because guessing is a reality of multiple-choice items. On the other hand, proponents in favor of the Rasch Model argue that not only is it impossible to estimate the *c* parameter accurately but also that guessing is really a characteristic of the examinee and not the item (see Lord, 1977, and Wright, 1977). Since both the Rasch Model and the 3PLM are used in practice, the two IRT models will both be further investigated in this dissertation.

Calibration Methods

Once an IRT model is chosen, item and ability parameters are to be calibrated based on the item responses. Two calibration methods are studied widely in horizontal equating literature: concurrent calibration and separate calibration.

Concurrent Calibration

For concurrent calibration, item and ability parameters are estimated for two forms at the same time, resulting in just one common scale. Concurrent calibration typically requires only one computer run. This linking is simultaneously established through common items. No further linking is required with concurrent calibration. An in-

depth treatment of the estimation procedure is given on Bock and Zimowski (1997) and Baker and Kim (2004).

Separate Calibration

Separate calibration requires multiple computer runs and results in each test form on its own separate scale, which requires some kind of transformation to establish linking relationship between the two forms through common items under a common-item nonequivalent groups design (cf. Kolen & Brennan, 2004).

Quite a few methods can be used to estimate the intercept and slope of the transformation, such as the mean/sigma method (Marco, 1977), the mean/mean method (Loyd & Hoover, 1980), and the test characteristic curve methods, i.e. the Haebara method (Haebara, 1980) and the Stocking-Lord method (Stocking & Lord, 1983). The mean/sigma method estimates the intercept and slope by making the means and standard deviations of the b-parameter the same on the common items. The mean/mean method estimates the intercept and slope through making *a* and *b* parameters the same on the common items. The Haebara method estimates the intercept and slope through minimizing the sum of the squared difference between the item characteristic curves for a given ability. The Stocking-Lord method estimates the intercept and slope through minimizing the squared difference between the test characteristic curves for a given ability.

Research has been conducted to compare the performance among these four methods (Baker & Al-Karni, 1991; Hanson & Beguin, 2002; Hung, Wu, &Chen, 1991;

Kim & Cohen, 1992; Kim & Kolen, 2007; Kim & Lee, 2004; Ogasawara, 2000, 2001b, 2001c; Way & Tang, 1991). The general conclusion is that the characteristic curve methods (i.e., the Haebara method and the Stocking-Lord method) provide more accurate transformation constants compared to the mean/sigma and the mean/mean methods. The Haebara method and the Stocking-Lord method provide similar estimates, and no systematic difference in accuracy has been found (Kim & Kolen, 2007; Way & Tang, 1991). However, Kolen and Brennan (2004, p. 172-173) argued that the Haebara method has several theoretical advantages over the Stocking-Lord method. Nevertheless the Stocking-Lord method is the dominant method and more widely used in practice. Therefore, the Stocking-Lord method was used in this dissertation.

Comparisons of Concurrent and Separate Calibrations

Studies comparing the two calibration methods typically show that the two methods tend to produce somewhat different results (Hanson & Beguin, 2002; Kim & Cohen, 1998; Kim & Kolen, 2007). Hanson and Beguin (2002) conducted a simulation study of separate versus concurrent item parameter estimations. Their findings indicated that concurrent estimation generally resulted in lower error and thus provided more accurate results than separate calibration when the IRT model assumptions are satisfied. However, Hanson and Beguin (2002) also argued that the results of the study were not sufficient to recommend completely avoiding separate estimation in favor of concurrent estimation.

Kim and Cohen (1998) compared three methods for developing a common metric under IRT: separate calibration through the Stocking-Lord method, concurrent calibration based on marginal maximum a posteriori estimation (using MULTILOG), and concurrent calibration based on marginal maximum likelihood estimation (using BILOG). The simulations were all based on data that fit the 3PLM. They pointed out that for smaller numbers of common items, linking through separate calibration yielded smaller root mean square differences for both item discrimination and difficulty parameters; for larger numbers of common items, the three methods yielded similar results. In addition, the studies conducted by Beguin, Hanson, and Glas (2000) and Beguin and Hanson (2001) indicated that separate estimation is more robust to the violation of the unidimesionality assumption compared with concurrent estimation. One advantage of separate calibration is that it facilitates examining item parameter estimates for the common items to identify outliers through plots (i.e., for items with estimates that do not appear to lie on a straight line, Kolen and Brennan, 2004). Kolen and Brennan (2004) also suggested that separate estimation using the test characteristic curve methods seems to be safest and that concurrent calibration could be used as an adjunct to the separate calibration.

In this study, since an equating chain is used to maintain an established vertical scale across multiple forms, the separate calibration method will be applied in the scale maintenance procedure. The Stocking-Lord method will be used to obtain transformation constants due to its wide usage in practice.

Examinee Sample Characteristics

In equating, two groups of respondents are generally involved, one administered the new form to be equated and one administered a previously equated form. Harris (1993) reviewed research conducted to investigate the impact of sample characteristics on equating results and concluded that the more alike the samples are the better the equating results. However, current guidelines do not reveal how different is too different or how much worse an equating is when the group of examinees taking the new form are not equivalent to the group of examinees administered the old form.

In a study investigating sample characteristics in an equating context, Stocking and Eignor (1986) concluded that differences in mean true ability can cause differences in the resulting equivalent mean ability estimates. Harris (1987) examined the size and the similarity of the nonequivalent groups in common item equating and found mixed results. In an IRT setting, Skaggs and Lissitz (1986) raised the question of how different sample characteristics can be before the samples are viewed as representative of separate populations rather than dissimilar samples from the same population. Angoff and Cowell (1985) noted that when the tests are dissimilar, using different subgroups for samples resulted in different equating conversions. Angoff and Cowell (1986) also examined the assumption that the equating of parallel forms is population independent.

By comparing IRT equating methods to traditional equating methods, Skaggs (1990b) mentioned that, with a common item design, similar samples, and a representative set of common items, all methods tended to provide similar results.

However, when a common item design was used and samples differed, none of the methods could assure adequate results. Marco, Peterson, and Stewart (1979) stated that the adequacy of equating worsened as samples became more divergent. They suggested that IRT methods might be preferable when samples differed. However, Skaggs (1990a) drew a different conclusion, that IRT methods were at least as sensitive as conventional methods to sample differences. Skaggs (1990b) also concluded that any of the equating methods was adversely affected when the common item design was used with groups of examinees that differed.

Cook, Eignor and Schmitt (1988) found that when using the common item design, the Levine unequally reliable equating method, the equipercentile method, and the IRT true score equating method were all sensitive to differences in group ability. They all yielded different results as a function of differences in group ability. Another study conducted by Cook, Eignor and Schmitt (1989) indicated that the theoretical basis for observed score equating would be more affected by differences between the samples than would true score equating.

Kolen and Brennan (2004) provided some rules of thumb regarding group differences, based on their extensive experience with the common-item nonequivalent groups design. First, mean differences between the two groups of approximately .1 or less standard deviation unit on the common items seem to cause few problems for any of the equating methods. Second, mean group differences of around .3 or more standard deviation units can result in substantial differences among methods, and differences

larger than .5 standard deviation units can be especially troublesome. In addition, ratios of group standard deviations on the common items of less than .8 or greater than 1.2 tend to be associated with substantial differences among methods. They suggested that differences in group standard deviations have the potential to lead to differences among methods that are at least as great as those caused by differences in means.

In a simulation study, Lin and Dorans (2011) investigated the population invariance of vertical linking. The subpopulations were defined by the specific curriculum emphases of the grades. The construct shift across tests of different grades, as well as some other test characteristics (e.g., item difficulty) and the ability distributions of subpopulations, were manipulated to examine the population invariance of vertical linking. The results indicated that when there is a construct shift across tests to be linked population invariance should not be assumed without further investigation of the characteristics of the tests and the populations that the linking functions would be applied to.

Power, Turhan, and Binici (2012) evaluated the population sensitivity of vertical scaling results for a state reading assessment spanning Grades 3-10 and a state mathematics test spanning Grades 3-8. Subpopulations considered in the study included males and females. The 3PLM was used to calibrate math and reading items and a common item design was used to construct the vertical scale. Results indicated more similar growth patterns across grades for gender subgroups for the mathematics assessment than for the reading assessments.

While they have been studied frequently in the equating literature, examinee sample characteristics have been virtually ignored in the maintenance of vertical scale literature. Therefore because of their importance for possible scale drift over multiple forms, in this study *examinee sample characteristics* have been treated as one of the factors assumed to have potential impact on the resulting scales using different combinations of scale maintenance approaches.

Construction of Vertical Scale

Even though this dissertation does not focus on how to develop a vertical scale, constructing a vertical scale is the first step in investigating the process of equating new forms to a vertical scale. It involves conceptual, technical, and implementation decisions (Harris, 2007). Much research has been conducted in vertical scaling, most of which focuses on definitions of growth, scaling design, statistical methods, type of scales, and so forth (Harris et al., 2003; Harris, 2007; Kolen, 2003; Kolen & Brennan, 2004; Proctor, 2008; Tong, 2005). Research suggests that vertical scaling is design-dependent (Harris, 1991 & 2007; Kolen & Brennan, 2004; Tong & Kolen, 2007), group-dependent (Harris & Hoover, 1987; Skaggs & Lissitz, 1988; Slinde & Linn, 1979; Tong & Kolen, 2008), and method-dependent (Harris, 2007; Kolen, 1981; Skaggs & Lissitz, 1986; Tong & Kolen, 2007). This section reviews research on different patterns of within-grade variability, choice of IRT model, and calibration methods in developing a vertical scale.

Patterns of Within-grade Variability

The primary reason for creating vertical scales is to measure learning across time. Without an understanding of the nature of growth, it is not possible to clearly evaluate whether a vertical scale is functioning as it should. The nature of growth is more of a conceptual issue than a measurement one, dealing with child development, psychology, and how the educational curriculum is implemented (Harris et al., 2004; Harris, 2007). It is the nature of growth and whether within-grade variance should increase, decrease, or remain constant that are the key issues in debate over scale shrinkage in vertical scaling literature (e.g., Becker & Forsyth, 1992; Camilli, 1988; Pommerich, & Thissen, 1998; Yen & Burket , 1997). Under different scaling methods, the pattern of within-grade scale variability may vary over grade. Harris (2007) stated:

> *One problem in trying to address the issue of defining growth is that test publishers rarely make the information explicit. It seems that most definitions are determined operationally, based on a combination of empirical data, the test development process, and preconceptions regarding the nature of growth. For example, a practitioner who believes within-grade variance should remain constant over grades might not develop test specifications or a data collection design with this in mind, but might reject scaling methods that resulted in large changes in within-grade variance over grades.*

Scholars have provided different perspectives on why scale shrinkage can occur. Yen (1986) noted that when a 3PLM is applied to tests covering a broad range of difficulty, the variance of ability scales often decreases over grades as the tests become more difficult. Yen argued that scale shrinkage is a function of dimensionality. Camilli (1988) proposed that IRT scale shrinkage depended upon the degree of mismatch between item difficulties and examinees' abilities. Camilli, Yamamoto, and Wang (1993)

argued further that scale shrinkage is not an inevitable result of IRT scaling, but rather of particular applications of IRT.

Studies have also revealed inconsistent results with regard to scale shrinkage. Becker and Forsyth (1992) found expanding variances as grade increased for both Thurstone scaling and IRT scaling. Yen and Burket (1997) found no evidence of IRT scale shrinkage, and their results even showed modest expansion for one simulated situation. William, Pommerich, and Thissen (1998) examined an IRT 3PLM scaling procedure and two Thurstone scaling procedures. Their findings indicated that scale expansion or shrinkage varies when different scaling procedures are conducted and that IRT scaling often produces similar results to other procedures when they are applied to the same data sets. In response to the lack of clarity and consensus, Harris (2004) recommended that researchers and practitioners developing vertical scales should make as explicit as possible their assumptions regarding the nature of growth and how these affect the development of the scale (Harris, 2004). Because of the need for more study of within-grade ability score variability and its effects on vertical scale development, this present study includes a planned factorial simulation of ability score variability.

Scaling Methods

Historically, three scaling methods have been used for vertical scaling: 1) Hieronymus scaling (Peterson, Kolen, & Hoover, 1989), 2) Thurstone scaling (1925), and 3) IRT scaling. The detailed procedures for these three scaling methods can be reviewed in the previous literature (e.g., Kolen, 2006; Kolen & Brennan, 2004; Peterson et al.,

1989). As mentioned, to be consistent with previous studies on scale maintenance, IRT scaling methods are used in this dissertation. The procedure of this scaling method is described in the chapter on methodology (Chapter III). When IRT is used to construct a vertical scale, the process is complicated. Decisions need to be made when implementing IRT for vertical scaling, such as choice of IRT models, computer programs, calibration methods, linking methods under separate calibration, and so on. IRT models require that the items measure the same construct(s) as well as satisfy the assumption of local independence. If the assumptions of the model are met, IRT can potentially provide item-free and person-free measurement. It is possibly due to the perceived properties of item-invariance and person-invariance that IRT scaling has been more and more utilized in constructing a vertical scale.

Choice of IRT Models

Early investigations of IRT models in vertical scaling focused on the Rasch Model. The findings regarding its application in vertical scaling are not entirely consistent. In numerous studies, the Rasch Model was found to be inadequate for constructing vertical scaling (e.g., Goulet, Linn, & Tatsuoka, 1975; Slinde & Linn, 1978, 1979; Loyd & Hoover, 1980, Kolen, 1981, & Holmes, 1982), because guessing and the violation of unidemensionality raised serious problems and thus resulted in a bad model fit. Other research found that the Rasch Model seemed to be adequate for constructing vertical scales (e.g., Gutanfsson, 1979; Guskey, 1981; Patience, 1981).

Slinde and Linn (1978) applied the Rasch Model to equate subtests of items on a mathematics achievement test by grouping items into three categories based on their difficulty parameter and examinees into three categories based on their achievement. They found that an examinee of medium ability had a better score on a more difficult test if the estimates were obtained from a high ability group and had a better score on an easier test if the estimates were obtained from a low ability group. Loyd and Hoover (1980) also used the Rasch model in vertical scaling settings. Their findings supported Slinde and Linn's (1978) results. They noted that the results tended to differ depending on how the scaling was conducted.

Other studies supported the utility of the Rasch Model for vertical scaling. Gustanfsson (1979) claimed that the findings of the previous studies about the inadequacy of the Rasch Model were a result of regression effects. He also claimed that there was no correlation between the item discrimination and difficulty parameter estimates and suggested that the Rasch model could be successfully implemented.

Studies regarding the application of the 3PLM in constructing a vertical scale focused on the comparisons of the 3PLM to the Rasch Model and other traditional methods (e.g., Marco, Petersen, & Stewart, 1983; Harris & Hoover, 1987; Skaggs & Lissitz, 1986). The findings were mixed as well. For example, Marco et al. (1983) examined the Rasch Model and the 3PLM under a variety of conditions and noted that the 3PLM resulted in promising empirical outcomes. Harris and Hoover (1987) examined the effectiveness of the 3PLM for vertical scaling, using data from a test of mathematics

computation. They noted that the 3PLM did not provide person-free item parameter estimation because different characteristic curves were obtained by test level and group. In a review of the research on horizontal equating and vertical scaling, Skaggs and Lissitz (1986) concluded that the 3PLM produced better results than the Rasch Model. They also noted that the calibration methods of transforming item parameters to the same scale can greatly affect the results based on the 3PLM. They suggested that if vertical scaling needs to be done, it would be safest to use either equipercentile methods or the 3PLM without concurrent calibration.

In general, the scholarship has provided mixed results and guidance with regard to choosing IRT models for horizontal equating and vertical scaling. Some scholarship found the Rasch Model to be acceptable. However, some scholarship found it to be unacceptable and still other scholarship has demonstrated better results with the 3PLM than the Rasch Model. Until the conditions and assumptions affecting vertical and horizontal scaling are more definitively clarified, test developers need to exercise care and provide careful oversight of results. This present study investigates factors affecting vertical scale quality in order to make a small contribution to increasing this needed clarity.

Concurrent and Separate Calibrations

Numerous studies have examined the relative effectiveness of concurrent calibration as compared to separate calibration. Several studies found that both methods produced similar results (Beguin & Hanson, 2000; Chin, Kim, & Nering, 2006; Hanson

& Beguin, 2006; Kim & Cohen, 1998). Other studies showed that the performance of the two methods varied based on specific conditions (Baker & Al-Karni, 1991; Jodoin, Keller, & Swaminathan, 2003; Karkee, Lewis, Hoskens, Yao, and Haug, 2003; Kim, Frisbie, & Kim, 2007; Meng, Wang, Viespol, Lee, & Wang, 2007; Kim, Lee, Kim, & Kelly, 2009).

The studies conducted by Beguin, Hanson, and Glas (2000) and Beguin and Hanson (2001) indicated that separate estimation is more robust to the violation of the unidimesionality assumption compared with concurrent estimation. Hanson and Beguin (2002) conducted a simulation study of separate versus concurrent item parameter estimations. The findings indicated that concurrent estimation generally resulted in lower error and thus provided more accurate results than separate calibration, when the parameter model assumptions are satisfied. Jodoin et al. (2003) compared separate calibration, fixed common item parameter estimation and concurrent calibration using data from three consecutive years. The findings suggested that the choice of the estimation procedure may have significant consequences in practice. Kim and Kolen (2005) study also supported this finding. Karkee et al. (2003) compared separate calibration, concurrent calibration, and a pair-wise concurrent calibration, using an elementary math test across six levels. They suggested that separate calibration produced the most consistent results among the three methods.

In addition to comparisons of separate and concurrent calibration, research has also compared the different methods of estimating linear linking functions in the process

of separate calibration (e.g., Baker & Alkarni, 1991; Hung, Wu, & Chen, 1999). In general, these studies concluded that the test characteristic curve methods for estimating linear linking functions tended to produce more accurate results.

In keeping with the findings from this literature, this study used separate calibration. The Stocking-Lord method will be used to obtain the transformation constants for separate calibration due to its wide usage in practice.

Proficiency Estimator

Through the calibration process, all item parameters are placed on the same scale and examinee proficiency is estimated. When IRT methods are used, θ can be estimated either using the entire response string for an examinee (referred to as pattern scoring) or by converting the number-correct (NC) scores (referred to as summed scoring) to the θ-scale. Four major classes of proficiency estimates can be obtained: *Maximum Likelihood Estimator* (MLE), *Expected A Posteriori* (EAP), *Maximum A Posteriori* (MAP), and Q*uadrature Distribution* (QD). MLK, EAP, and MAP estimates provide individual proficiency estimates, whereas QD estimates the entire posterior distribution and does not provide individual proficiency estimates. Lord (1980) and Baker (1992) described various estimation methods. Thissen and Orlando (2001) gave an extended discussion for proficiency estimation.

MLE estimates are obtained by maximizing the likelihood function of the item parameters and a given string of responses. For long tests, MLE is an unbiased estimate of proficiency, but the MLE does not exist for scores of all correct or zero. The computer

program used for estimation will force these scores to take on either extremely large or small values to aid in convergence. EAP and MAP are obtained using Bayesian procedures. The EAP estimator is the mean of the posterior distribution, and the MAP estimator is the mode of the posterior distribution (Thissen& Orlando, 2001). Both of these estimators will result in biased estimates of proficiency, but both will provide smaller variability of estimates than MLE (Lord, 1986).

Estimates using *Maximum Likelihood Estimator* (MLE) based on pattern scoring have been commonly used. Research has investigated the other three proficiency estimators using pattern scoring (Hendrickson, Kolen, & Tong, 2004; Kim, Frisbie, Kolen, & Kim, 2007; Kim, Lee, Kim, & Kelly, 2009; Meng, Kolen, & Lohman, 2006; Tong, 2005). Other research also has involved proficiency estimators using summed scoring (Kim, Lee, Kim, & Kelly, 2009; Proctor, 2008; Thissen & Orlando, 2001; Tong, 2005; Yen, 1984b).

Summary

Previous scholarship suggests that maintaining vertical scales across forms is an involved process with many decisions. Many potential factors in the scale maintenance process may affect the resulting scale. The factors include scale maintenance approaches, within-grade variability patterns, and examinee sample characteristics. A different decision on any of the factors listed above may lead to a somewhat different resulting scale using different approaches to equating new forms to an established vertical scale. In addition, the literature about constructing a vertical scale suggests that maintaining the

vertical scale across forms is also sensitive to how a baseline vertical scale is established. However, how to maintain a vertical scale across multiple forms has not been well addressed in the existing literature and needs to be further considered.

Comparisons of the scale maintenance approaches among the research studies in the literature indicate that scale maintenance approaches often yield similar results over a second form, at least in the conditions examined in the literature to date. It is possible that there will be more drift over a chain of multiple forms. Changes over more than two forms need to be investigated, as drift may be small and not noticeable until a longer chain is examined. Because it has been inadequately studied in the previous literature, linking over three forms in vertical scaling is the central focus in this study. The purposes of the dissertation are to compare how different the resulting proficiency estimates are by using two scale maintenance approaches, and thus to investigate under which conditions of within-grade variability patterns and examinee sample characteristics one approach is preferable to the other given an established vertical scale. The next chapter outlines the research questions and the methodology used to answer them.

CHAPTER III
METHODOLOGY

When new forms of an assessment are developed, an equating design must be incorporated as part of the implementation of the new forms. Hoskens, Lewis, and Patz (2003) noted that a form-to-form vertical scale equating design should foster scale comparability of the scale across grades within form, within grade across forms, and across grades and across forms. Recent studies comparing different approaches for maintaining a vertical scale over two forms have yielded mixed results in the conditions examined (cf. Cao et al., 2007; Hoskens et al., 2003; Tong & Kolen, 2008; Tong & Kolen, 2009; Wang & Harris, 2009a; Wang & Harris, 2009b; Tomkowicz, Zhang, & Yen, 2010). No consensus exists as to which scale maintenance approach is preferable in any given situation over two forms. In addition, more drift over a chain of multiple forms is a distinct possibility. Therefore, linking over three forms in vertical scaling is the central theme of this dissertation, as drift may be small and not noticeable until a longer chain is examined. In summary, the purposes of this dissertation are to compare how different the resulting proficiency estimates are by using two different scale maintenance approaches in supporting scale stability across grades within form, within grade across forms, and across grades and across forms, and thus to investigate under which conditions of within-grade variability patterns and examinee sample characteristics one approach is preferable

to the other. The similarity and dissimilarity of the resulting proficiency estimates created by these two approaches is the focus of this dissertation.

As mentioned in Chapter I, six conditions were investigated in this study, considering three different within-grade variability patterns and two levels of examinee group differences. Under each condition, three test forms were developed, each consisting of three level tests through Grade 10 to Grade 12. The procedures resulted in six sets of data combinations organized in the six conditions.

Three criteria – correlation, Root Mean Square Error, and Mean Absolute Difference – were used to examine how adequate the two scale maintenance approaches are in recovering the true examinee proficiencies, under the six conditions of proficiency distributions. To capture the growth trends established by the six conditions, three properties proposed by Kolen and Brennan (2004) were used as criteria: grade-to-grade growth, grade-to-grade variability, and separation of grade distributions. To compare the performance of the two scale maintenance approaches under different conditions of within-grade variability pattern and examinee sample characteristics, the ratios of RMSEs and the ratios of MADs between the vertical approach and the horizontal approach within form were used, as well as the ratios of RMSEs the ratios of MADs between the new Form X and the interim Form Z via each approach.

The five sections of this chapter describe (a) the test form development procedure, (b) the steps of data generation, (c) conditions for constructing vertical scales, (d) scale maintenance approaches, and (e) evaluation criteria.

Test Form Development

Linking across three forms in vertical scaling involves a linking through which the new forms are linked to an established vertical scale. In this dissertation, the three forms were denoted as Form Y (baseline form), Form Z (interim form), and Form X (new form). Each form contained three level tests, one for each grade level 10-12. The scales on Form X were transformed to the established vertical scale on Form Y through the interim Form Z (X to Z to Y). For constructing the vertical scale, a baseline vertical scale was established spanning from Grade 10 to Grade 12 on the baseline form (Form Y). This study linked across three forms to a baseline vertical scale in a common-item nonequivalent groups design. There were horizontal common items for the same grade level across Form Y and Form Z, Form Z and Form X, and vertical common items between adjacent grades within the same form. For instance, for the Grade 10 test across forms, Form Z shared horizontal common items both with Form Y and with Form X. Within a given grade there was no overlap between the two sets of horizontal common items. Similarly, there were vertical common items between Grade 10 and Grade 11 for a given form. The horizontal common items for the same grade between forms were used to horizontally equate the tests via a linking chain: Form X to Form Z to Form Y. The vertical common items between two adjacent grades for a given form were used to establish a linking relationship spanning all grade levels.

Since the common-item design was used for both vertical scale construction and scale maintenance in this dissertation, the level tests on the three forms to be developed

should contain items in common not only between Form Y and Form Z and Form Z and Form X (horizontal common items) but also between the grade and its adjacent ones within form (vertical common items). In this dissertation, the vertical common items between any adjacent grades included both items shared with the higher of the two grades and items shared with the lower of the two grades. In addition, the horizontal common items were targeted to be representative of the overall level test in terms of content representation and range of difficulty. There was no overlap between the vertical common items and the horizontal common items within each level test, which means none of the vertical common items and the horizontal common items were the same within each level test. Table 3.1 shows the item blocks for the level tests in the three test forms. For each of the three test forms, there are eight item blocks across three grade levels (Grades 10 - 12). Each level test, accordingly, consists of blocks of items, which are unique for that grade level, or vertical common blocks, or horizontal common blocks. For example, in addition to one or two vertical-common-item blocks between two adjacent grades, each of the level tests on the interim form (Form Z) contains two blocks of horizontal common items, one in common with Form Y and the other with Form X.

The level tests for each form were developed on the basis of the content specification of the ACT English Test. The ACT English Test consists of 75 multiple-choice items. The English test measures six content areas: punctuation, grammar and usage, sentence structure, strategy, organization, and style. Table 3.2 shows the item proportion and corresponding number of items within each content area on the ACT

English Test. In this dissertation, the level tests for each form were designed to measure students' progressive development of knowledge and skill in the same six content areas. Thus, all the level tests had similar content specifications and similar item proportions as ACT English Test. The level test for Grade 10 was intended be shorter and less difficult than that for Grade 11, and the level test for Grade 11 is also intended to be shorter and less difficult than that for Grade 12. Forty easy items were chosen for the Grade 10 test, forty-five items of median difficulty for the Grade 11 test, and forty-seven of the most difficult items for the Grade 12 test. Table 3.3 shows the corresponding numbers of items within content area for each level test, similar to the item proportions of the 75-item ACT English Test.

This dissertation used 375 items from five 75-item ACT English forms denoted Form A, B, C, D, and E, as the base item pool of simulating data to mock up three grade level tests (Grades 10-12) for Forms Y, Z and X. The five original tests were all administered nationally. Since the data were used to generate level tests through Grades 10-12, the sample sizes for these three grades were very large. For example, the samples of Grades 10-12 students who took Test Form A were about 5000 from Grade 10, about 60,000 from Grade 11, and about 300,000 from Grade 12. In order to make the level tests simulated for this dissertation more realistic to operational tests, the level tests in Forms Y, Z, and X were developed based on similar match-to-specification criteria as the ACT English Test. Form Z as the interim was created first following the match-to-specification criteria, then Form Y as the old form, and finally Form X as the new form. As shown in

Table 3.1, Form Z consists of two blocks of horizontal common items at the same grade, one common with the old Form Y and the other common with the new Form X. These blocks of items were used to develop linking relationships through an equating chain. Table 3.4 lists the number of items within content area in each item block of the level tests in Form Z. For example, the horizontal item block 1 of Grade 10 level test, denoted as *HC1_10*, consists of fifteen items, with two in punctuation, three in grammar and usage, three in sentence structure, three in strategy, two in organization, and two in style. The number in the parenthesis is the total number of items of each item block. Using the same algorithm as Form Z, Form Y and Form X were constructed accordingly.

All the 375 items in the base item pool were calibrated through concurrent calibration, to ensure the item parameters used in the data simulation were on the same scale. BILOG-MG 3 (Zimowski, Muraki, Mislevy, and Bock, 1996) was used to execute the IRT 3PLM item parameter estimations, and proficiency estimates were obtained via maximum-likelihood estimator (MLE). Items for each level test were selected based on the values of item difficulty parameter estimates from concurrent calibration. The level test for Grade 10 was intentionally shorter and less difficult than that for Grade 11, and the level test for Grade 11 was intentionally shorter and less difficult than that for Grade 12. Forty easy items were chosen for the Grade 10 test, forty-five items of median difficulty for the Grade 11 test, and forty-seven of the most difficult items for the Grade 12 test. Once the level tests were constructed, the corresponding item parameter estimates of the selected items were used as true item parameters for the data simulation. Table 3.5

lists the average item parameter estimates of each level test with Forms Y, Z, and X, as well as the corresponding average p-values and point bi-serials.

Data Generation

Since the issue of linking across forms in vertical scaling is the central theme of this dissertation, item responses for a pair of original and new forms were generated simultaneously. As mentioned above, items from the five forms of the ACT English Test were concurrently calibrated using BILOG-MG assuming a 3PLM. These estimated item parameters were treated as population item parameters for simulating data.

Generating Proficiency Parameters

The proficiency distributions of θ-parameter estimates for each grade based on the real data were used to generate θ-parameters. Two factors, the pattern of within-grade proficiency distributions across grades and examinee sample characteristics, were considered in generating proficiency parameters.

Pattern of Within-grade Proficiency Distribution across Grades

On the basis of the first factor, three conditions were considered for each of the three forms. In all three conditions, all three grade groups have a normal distribution of proficiency parameters. The three conditions in the three forms are as follows:

1. The mean of θ-parameters for each grade is the same as the corresponding mean estimate based on the real data, but the SD for each grade is held to be 1.

2. The mean of θ-parameter for each grade is the same as the corresponding mean estimate, and the SD for each grade is decreasing from Grade 10 to Grade 12 based on the real data.
3. The mean of θ-parameters for each grade is the same as the corresponding mean estimate based on the real data, but the SD for each grade is specified to be increasing from Grade 10 to Grade 12.

Examinee Sample Characteristics

The proficiencies of the three groups were assumed to be normally distributed. The proficiencies of Group 3 taking Form X were generated by assuming the proficiencies were higher than those for Group 2 taking Form Z; the proficiencies of Group 1 taking Form Y were generated by assuming them to be lower than those for Group 2 taking Form Z. Simulation studies under common-item nonequivalent groups designs in the literature used different values of mean proficiency difference between the two populations, among which three values were often used: 0.2 (Nozawa, 2008; Tong, 2005), 0.25 (Wang, Kolen, & Harris, 2000; Wang, Lee, & Brennan, 2006; Wang, Lee, & Brannan, 2009), and 0.5 (Cao et al., 2007; Hanson & Beguin, 2002; Park, Young, & Yi, 2008). Examinee sample characteristics may differ in various ways, yet it is beyond the scope of this dissertation to examine all kinds and levels of variability. Therefore, two values of mean proficiency difference -- 0.25 and 0.5 -- were taken into consideration in this dissertation when generating proficiency parameters.

1. The mean of each grade-level proficiency distribution on Form X was increased by 0.25 compared with those on Form Z; the mean of each grade-level proficiency distribution on Form Z was increased by 0.25 compared with those on Form Y. Thus the mean proficiency difference between Forms X and Y within each grade is 0.5.
2. The mean of each grade-level proficiency distribution on Form X was increased by 0.5 compared with those on Form Z; the mean of each grade-level proficiency distribution on Form Z was increased by 0.5 compared with those on Form Y. Thus the mean proficiency difference between Forms X and Y within each grade is 1.

These two factors (3 x 2) are fully crossed and thus 6 conditions were considered when generating ability parameters. Table 3.6 shows the six conditions of true proficiency distribution for the simulations. The means and SDs used to generate these distributions under the six conditions are listed in Table 3.7. Under the six conditions, three sets of vertical scales were established for the base Form Y, using three patterns of within-grade variability. To generate data for Forms Z and X, the examinee sample characteristics – small group difference and large group difference – were taken into consideration.

Calculating the Probability

Based on the generated item and ability parameters, the probability of a correct response was calculated for each combination of examinees and items. If a random

number generated from the standard uniform distribution was smaller than the probability of a correct response, the corresponding item response variable took on a value of 1. Otherwise, the item response variable took on a value of 0.

Six sets of data were generated, each of which consists of a pair of item responses (one for the original form and the other for the new forms). For each of the six simulation conditions, 100 sets of data were generated for replication with a sample size of 2000 per grade for each form.

Construction of Vertical Scales

After data generation, the next step was to construct a vertical scale for the original form (Form Y). The IRT scaling variability at this stage involved three conditions of patterns of within-grade proficiency variability (increasing, decreasing, and constant), and thus 3 possible sets of vertical scale were established based on Form Y for scale maintenance investigation.

To obtain item parameter estimates based on the 3PLM, separate calibrations for each grade for each year's data were conducted using BILOG-MG 3 (Zimowski, Muraki, Mislevy, & Bock, 1996). In separate calibrations, the vertical common items were used to place item parameter estimates, examinee ability estimates and estimated ability distributions on the base grade scale using the Stocking-Lord (1983) characteristic curve transformation method. The computer program *ST* (Zeng & Hanson, 1995; Cui, 2004) was run to compute the estimates for slopes and intercepts for the estimated latent

distributions. A chaining process was conducted to link estimates for the grades that are not adjacent to the base grade (See details in Kolen & Brennan, 2004).

Based on the above description, the following steps were used when separate calibrations were applied for the baseline form:

1. Calibrate the level tests separately, one grade at a time;
2. Estimate the slopes and intercepts through vertical common items using the Stocking-Lord approach;
3. Place all estimates onto the same vertical scale;
4. Obtain ability distribution of MLE estimation.

Grade 10 was first scaled to have a mean of 0 and an SD of 1, and the other grades were scaled accordingly. Through a linking chain, a vertical scale was established, spanning Grade 10 to Grade 12. The computer program *ST* (Hanson & Zeng, 2004) was used to estimate the Stocking-Lord transformation constants, using item parameter estimates for the vertical common items from two adjacent grades.

Scale Maintenance

Given that a vertical scale was established for Form Y, there were two possible ways to maintain the vertical scale on Form X through the interim Form Z: *constructing a base vertical scale and maintaining it through horizontal equating* and *constructing separate vertical scales for each of the three forms and horizontally linking the vertical scales*. For scale maintenance, two scale maintenance approaches, the horizontal

approach and the vertical approach, were investigated, given each of the 3 sets of vertical scales established on the base Form Y.

Horizontal Approach

Under this approach, vertical linking items were needed on Form Y to establish the vertical scale. No vertical linking items were needed on Form Z and Form X. The two sets of horizontal common items for the same grade across forms were used to obtain corresponding transformation constants between Forms Y and Z and between Forms Z and X. Through an equating chain, the scale on Form X were placed onto Form Y's vertical scale.

Separate calibration was applied in scale transformation. Item parameter estimates, ability estimates and ability distributions of the horizontal common items for each grade across the three forms were obtained through separate calibration using BILOG-MG 3. There were four major steps under the horizontal approach after separate calibration:

1. Establish a baseline vertical scale on Form Y using vertical common items between adjacent grades;
2. Obtain the transformation constants *S* and *I* between Form Z and Form X via the Stocking-Lord method using the horizontal common items between Form Z and Form X;

3. Obtain the transformation constants between Form Z and Form Y via the Stocking-Lord method using the horizontal common items between Form Z and Form Y;
4. Use Form Y as the base form and place the rest of the forms onto the vertically adjusted scale on Form Y within each grade, using an equating chain.

The equating chain was used to link parameter estimates for Form X onto the vertical adjusted scale on Form Y within a given grade. For example, to link Form X to Form Y within Grade 10, the scale on Form X was linked to the scale on Form Z first using the Stocking-Lord method obtained on the horizontal common items between Form Z and Form X. Next, the results from the previous step were scaled onto the vertically adjusted scale on Form Y through the Stocking-Lord method obtained based on the horizontal common items between Form Z and Form Y. Thus a total of two transformations were involved to place Form Z onto the scale on Form Y.

Vertical Approach

Under this approach, three separate vertical scales were developed using respective vertical common items, one on Form Y, one on Form Z, and the other on Form X. Next, the vertical scale developed on Form X was linked back to the vertical scale on Form Y via the interim Form Z through an equating chain.

To achieve this linking of the two vertical scales on Form X and Form Y via Form Z, horizontal common items were used to identify the linking relationship. They are the same two sets of horizontal common items as those used in the horizontal approach;

under the vertical approach, these items were used to link vertical scale, instead of linking tests within the same grade. To obtain the corresponding linking transformation constants between the vertical scales on Forms Y and Z and between the vertical scales on Forms Z and X, the item parameters for each grade level on all three forms were placed onto their respective vertical scales. The transformation constants could not be used directly from the previous horizontal approach because those constants were on the same scale as each of the grades but not on the vertical scale.

Separate calibrations were used via BILOG-MG 3 respectively under the 3PLM. The procedure for separate calibrations was the same as that for the previous horizontal approach. Because there were common items for each of the three grades between Form Y and Form Z, for example, potentially three sets of transformation constants could be computed through separate calibration. Which set should be applied to establish linking? If the transformation constants for all three grades were the same or similar, it would not matter much which grade was used to establish the horizontal linking of the two vertical scales. However, if the transformation constants across grades were too different from each other, the transformation constants chosen would impact the results. One way to proceed is to average all the equating constants and to use the average to place item parameter estimates of horizontal items on the same scale. An alternative would be to obtain a single set of transformation constants based on all the common items. Since this alternative is probably more justifiable than averaging constants, this approach was used.

In summary, there are four major steps under this approach:

1. Develop a vertical scale for each of the three forms: Form Y, Form Z and Form X;
2. For each form, place item parameter estimates for all grades onto the developed vertical scale using vertical common items;
3. Compute the single set of transformation constants through the Stocking-Lord method based on the vertically scaled item parameter estimates for all the horizontal common items between Forms Y and Z across all grade levels;
4. Compute the single set of transformation constants through the Stocking-Lord method based on the vertically scaled item parameter estimates for all the horizontal common items between Forms Z and X across all grade levels;
5. Link the vertical scale on Form X to the vertical scale on Form Z using the transformation constants obtained in step 4;
6. Link the adjusted vertical scale on Form X from step 4 to the vertical scale on Form Y, using the transformation constants obtained in step 3.

Evaluation Criteria

A strong advantage of conducting this simulation study is that objective criteria exist to evaluate the results since true item parameters and student proficiencies are known. Two vertical scale maintenance approaches were introduced and compared to examine their practical impact on the resulting proficiency estimates, in combination with model choices and examinee sample characteristics. Three research questions guided the study:

1. How adequate are the two scale maintenance approaches in recovering the true examinee proficiencies, under the six conditions of proficiency distributions, considering the factors of within-grade variability patterns and examinee sample characteristics?
2. What are the effects of the two scale maintenance approaches on the resulting proficiency estimates and growth interpretations, under the six conditions of proficiency distributions, considering the factors of within-grade variability patterns and examinee sample characteristics?
3. Under which conditions of proficiency distributions does one scale maintenance approach provide more equivalent resulting proficiency estimates than the other, considering the factors of within-grade variability patterns and examinee sample characteristics?

To answer these three questions, the following evaluation criteria were applied.

Adequacy in Recovering the True Proficiency

With regard to the first research question of this dissertation -- how adequate the different combinations of scale maintenance approaches are in recovering the true examinee proficiencies -- three criteria were used to evaluate the overall fit of the different combinations of scale maintenance approaches, within-grade variability patterns, and examinee sample characteristics, given an established vertical scale.

The first criterion is the correlation between the true proficiencies and the estimated proficiencies of the examinees, defined as the covariance of the true proficiency parameter and the proficiency estimate divided by the product of their standard

deviations. The correlation used is the Pearson (1896) product-moment coefficient, which is expressed as

$$r_{\theta,\hat{\theta}} = \frac{\text{cov}(\theta, \hat{\theta})}{\sigma_{\theta}\sigma_{\hat{\theta}}} \tag{3}$$

where θ represents the true proficiency of an examinee, and $\hat{\theta}$ represents the estimated proficiency of the examinee. The average of the correlations is taken over the 100 replications. A higher correlation average indicates that the scale maintenance approach is a better fit for the data.

The Root Mean Square Error (RMSE) and the Mean Absolute Difference (MAD) are used as the second and third criteria, showing the extent to which the estimated proficiency values match the true values using different scale maintenance approaches to link across forms in vertical scaling. The RMSE and the MAD are expressed as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}{N}} \tag{4}$$

$$MAD = \frac{\sum_{i=1}^{N}\left|\hat{\theta}_i - \theta_i\right|}{N} \tag{5}$$

where *N* is the sample size of 2000, $\hat{\theta}_i$ represents the proficiency estimates of examinee *i*, and θ_i represents the true proficiency value of examinee *i*. The smaller the RMSE, the better the scale maintenance approaches is. Since the errors are squared before they are averaged as shown in Formula 4, the RMSE gives a relatively high weight to large error. The MAD in Formula 5 is a linear score which means that all the individual differences are weighted equally in the average. The RMSE will always be larger or equal to the

MAD. The RMSE index was chosen for this study because it is widely used as a criterion in the equating studies to compare the results using different equating methods. Mathematically speaking, the RMSE differences between the two scale approaches should be similar to the MAD differences between the two scale approaches, even though the magnitudes of MADs are smaller than the RMSEs. Therefore, the MAD index is used in this study as an adjunct index to check whether any inconsistency exists while comparing the MAD differences with the RMSE differences between the two approaches. The average of the RMSE is taken over 100 replications. The average of the MAD is taken over 100 replications as well.

Growth Trend

With regard to the second research question – the effect of the combinations of scale maintenance approaches on the resulting scale and growth interpretation -- three properties proposed by Kolen and Brennan (2004) were employed: grade-to-grade growth, grade-to-grade variability, and separation of grade distributions. These three properties were used to capture the growth trends established by the two scale maintenance approaches under the six conditions.

For the grade-to-grade growth, the mean proficiency difference between adjacent grades was calculated to check the overall growth of the students from lower grade to higher grade. The pattern of grade-to-grade growth was represented by means and mean differences between adjacent grades. For the grade-to-grade variability, the standard deviations of the scale scores between the adjacent grades were calculated. The pattern of grade-to-grade variability is represented by standard deviations within and across grade levels. The third property, the separation of grade distributions, is examined using an

index proposed by Yen (1986) for the effect size of grade-to-grade differences. In vertical scaling context, the effect size is defined as

$$effectsize = \frac{\bar{\hat{\theta}}_{upper} - \bar{\hat{\theta}}_{lower}}{\sqrt{(\hat{\sigma}^2_{upper} + \hat{\sigma}^2_{lower})/2}} \quad (6)$$

where $\bar{\hat{\theta}}_{upper}$ and $\bar{\hat{\theta}}_{lower}$ represents the mean proficiency estimates for the higher and lower grades, and $\hat{\sigma}^2_{upper}$ and $\hat{\sigma}^2_{lower}$ represent the estimated proficiency variances for the upper and lower grades.

Scale Maintenance Approach Preference

With regard to the third research question, under which condition of within-grade variability patterns and examinee sample characteristics one approach is preferable to the other, comparative ratios were computed for each of the six conditions of within-grade variability patterns and examinee group differences. Four ratio indices of the average RMSE across Grades 10-12are calculated separately under the six conditions:

1. the ratio of the average RMSE for the vertical approach divided by the average RMSE for the horizontal approach on the interim Form Z;
2. the ratio of the average RMSE for the vertical approach divided by the average RMSE for the horizontal approach on the new Form X;
3. the ratio of the average RMSE for the horizontal approach on the new Form X divided by the average RMSE for the horizontal approach on the interim Form Z;

4. the ratio of the average RMSE for the vertical approach divided by the average RMSE for the horizontal approach.

In addition, four similar ratio indices of the average MAD were computed. Thus, twenty-four ratio indices were computed, with eight ratio indices under each of the six conditions. In short, instead of comparing the estimated proficiency values with their true values, these ratios summarized the relative discrepancies in the error introduced by using the two different scale maintenance approaches on the interim Form Z and on the new Form X, under the six specified conditions.

In sum, evaluation criteria were identified to answer the three research questions of this dissertation.

To address the first research question, Pearson product-moment coefficient, RMSE, and MAD are the three indices used to investigate how adequate the different combinations of scale maintenance approaches are in recovering the true examinee proficiencies.

To investigate the effect of difference combinations on the resulting proficiency estimates, three properties were used: patterns of grade-to-grade growth, patterns of grade-to-grade variability, and separation of distributions. The indices -- mean, mean difference between adjacent grades, SD, and effect size -- were computed for all the estimated proficiency distributions obtained from different combinations of scale maintenance approaches. The results were compared to address the second research question.

To address the third research question, under which combination one approach is preferable to the other, the RMSE ratios and the MAD ratios between the two scale maintenance approaches were computed to show the discrepancies in the error introduced by the two scale maintenance approaches in the process of linking the proficiency scales across forms under the different specified conditions. These comparisons help decide which the scale maintenance approach is preferable to the other under each of the six conditions.

Table 3.1 Blocks of Items for the Level Tests

Test Form	G10	G10/11	G11	G11/12	G12
Form Y	Unique(15) HC[a] (15)	VC[c] (10)	Unique(12) HC[a] (12)	VC (11)	Unique(18) HC[a] (18)
Form Z	HC[a](15) HC[b] (15)	VC (10)	HC[a](12) HC[b] (12)	VC (11)	HC[a](18) HC[b] (18)
Form X	HC[b] (15) Unique(15)	VC (10)	HC[b] (12) Unique(12)	VC (11)	HC[b] (18) Unique(18)

Note: HC[a]: Horizontal Common Items between Form Y and Form Z for a given grade.
HC[b]: Horizontal Common Items between Form Z and Form X for a given grade.
VC[c]: Vertical common items between two adjacent grades within form.

Table 3.2 Content Specification, Item Proportion, and Number of Items on ACT English Assessment Test

	ACT English Assessment Test	
Content	Item Proportion	Number of Items
Punctuation	0.13	10 +/- 1
Grammar & usage	0.16	12 +/- 1
Sentence structure	0.24	18 +/- 1
Strategy	0.16	12 +/- 1
Organization	0.15	11 +/- 1
Style	0.16	11 +/- 1
Total		75

Table 3.3 Numbers of Items for the Three Level Tests through Grades 10-12

Content	Item Proportion	G10 Number of Items	G11 Number of Items	G12 Number of Items
Punctuation	0.14	6 +/- 1	6 +/- 1	6 +/- 1
Grammar & usage	0.18	7 +/- 1	8 +/- 1	8 +/- 1
Sentence structure	0.22	11 +/- 1	12 +/- 1	12 +/- 1
Strategy	0.17	5 +/- 1	6 +/- 1	6 +/- 1
Organization	0.15	5 +/- 1	6 +/- 1	9 +/- 1
Style	0.14	6 +/- 1	6 +/- 1	7 +/- 1
Total		40	45	47

Table 3.4 Number of Items by Content Area for Item Blocks of the Level Tests Grades 10-12 Form Z

	G10			G11			G12	
Content	HC1_10 (15)	HC2_10 (15)	VC_10-11 (10)	HC1_11 (12)	HC2_11 (12)	VC_11-12 (11)	HC1_12 (18)	HC2_12 (18)
PUN	2	2	1	2	2	1	3	3
BGU	3	3	2	2	2	2	3	3
SST	3	3	2	2	2	2	4	4
STR	3	3	2	2	2	2	3	3
ORG	2	2	2	2	2	2	3	3
STY	2	2	1	2	2	2	2	2
Total	15	15	10	12	12	11	18	18

Table 3.5 Level Test Average Item Parameter Estimates, Average Classic Difficulty, and Average Discrimination Values in Forms Y, Z, and X

		Grade 10	Grade 11	Grade 12
Form Y	irt_a	0.723	0.808	0.943
	irt_b	-0.258	0.336	0.975
	irt_c	0.159	0.168	0.198
	p-value	0.630	0.580	0.530
	pbsr	0.500	0.520	0.540
Form Z	irt_a	0.709	0.750	0.883
	irt_b	-0.307	0.335	0.935
	irt_c	0.186	0.155	0.181
	p-value	0.630	0.570	0.510
	pbsr	0.500	0.530	0.540
Form X	irt_a	0.702	0.714	0.872
	irt_b	-0.300	0.333	0.930
	irt_c	0.188	0.163	0.200
	p-value	0.630	0.570	0.530
	pbsr	0.490	0.500	0.500

Table 3.6 Six Conditions of Simulated Proficiency Distributions

	Examinee Sample Characteristics	
	Small group difference	Large group difference
Patterns of Within-grade Variability	(d_{mean} = 0.25)	(d_{mean} = 0.50)
Constant SD	Condition 1	Condition 2
Decreasing SD	Condition 3	Condition 4
Increasing SD	Condition 5	Condition 6

Note: d_{mean} refers to the mean proficiency difference between two samples taking different forms.

Table 3.7 Means and SDs for the Simulated Proficiency Distributions under Six Conditions

	Form Y		Form Z		Form X	
	Mean	SD	Mean	SD	Mean	SD
Condition1						
G10	0.000	1.000	0.250	1.000	0.500	1.000
G11	0.666	1.000	0.916	1.000	1.166	1.000
G12	1.022	1.000	1.272	1.000	1.522	1.000
Condition 2						
G10	0.000	1.000	0.500	1.000	1.000	1.000
G11	0.666	1.000	1.166	1.000	1.666	1.000
G12	1.022	1.000	1.522	1.000	2.022	1.000
Condition 3						
G10	0.000	1.000	0.250	1.000	0.500	1.000
G11	0.666	0.921	0.916	0.921	1.166	0.921
G12	1.022	0.631	1.272	0.631	1.522	0.631
Condition 4						
G10	0.000	1.000	0.500	1.000	1.000	1.000
G11	0.666	0.921	1.166	0.921	1.666	0.921
G12	1.022	0.631	1.522	0.631	2.022	0.631
Condition 5						
G10	0.000	1.000	0.250	1.000	0.500	1.000
G11	0.666	1.200	0.916	1.200	1.166	1.200
G12	1.022	1.400	1.272	1.400	1.522	1.400
Condition 6						
G10	0.000	1.000	0.500	1.000	1.000	1.000
G11	0.666	1.200	1.166	1.200	1.666	1.200
G12	1.022	1.400	1.522	1.400	2.022	1.400

CHAPTER IV
RESULTS

This chapter presents the findings of the analyses outlined in Chapter III. The two scale maintenance approaches, horizontal linking and vertical linking, were conducted for linking simulated data across three forms, Forms Y, Z, and X, each of which consists of three level tests. Two factors, within-grade variability patterns and examinee sample characteristics, were considered in generating proficiency distributions. Since there is no universally agreed upon growth model in the literature, the three within-grade variability patterns with two group differences were all investigated, and thus six sets of data were generated accordingly.

The first section of this chapter describes the data generation process for the simulated data of six conditions of proficiency distributions. To address the research questions proposed in Chapter I, the following sections then report the resulting proficiency distributions and the related indices under each of the three pairs of proficiency distributions: Conditions 1 and 2, Conditions 3 and 4, and Conditions 5 and 6. Comparisons between the resulting proficiency estimates developed using the two scale maintenance approaches were made under the three pairs of proficiency distribution conditions.

Simulation Data Analysis

To investigate linking over forms in vertical scaling, three forms were developed as shown in Table 3.1, each of which contains three level tests. Thus nine level tests were developed altogether across Forms Y, Z, and X. As described in Chapter III, six sets of proficiency distribution combinations were generated from normal distributions with specific means and SDs. Table 3.6 describes the six conditions of proficiency distributions by considering two factors, within-grade variability patterns and examinee sample characteristics. The within-grade variability indices were specified in such a way that the underlying proficiency scale remained constant in the within-grade variability (Conditions 1 and 2), decreased (Conditions 3 and 4), and increased (Conditions 5 and 6). In addition, two values of the mean proficiency difference among the population across Forms Y, Z, and X were used (0.25 and 0.5) indicating small group difference (Conditions 1, 3, and 5) and large group difference (Conditions 2, 4, and 6). Table 3.7 lists the means and SD for the level tests of the three forms under each of the six conditions. The mean is always increased as grade increased, as is expected for a true development scale. The IRT scaling variability involved three within-grade proficiency variability conditions (constant, decreasing, and increasing), and thus three sets of base vertical scale were established on the baseline Form Y for scale maintenance investigation: one for Conditions 1 and 2, one for Conditions 3 and 4, and one for Conditions 5 and 6.

A sample size of 2000 examinees was used to generate data for each of the nine level tests under each of the six conditions. One hundred replications were run for each condition. These proficiency distributions were generated using the random generator in the computer software R (CTRAN, 2008). The means and SDs for the proficiency distributions of the 2000 simulated examinees for each level test across forms, over 100 replications, are reported in Table 4.1. As expected, these values are quite similar to the parameter values (see Table 3.7 in Chapter III) used to generate the proficiency distributions.

Item parameter and Data Simulation

After obtaining the proficiency distributions with the variability trends specified in Table 3.8, item responses were simulated simultaneously for the nine level tests across the three forms, Forms Y, Z, and X. This dissertation used 375 items from five 75-item ACT English forms as the base item pool of the simulated data to mock up level tests for Grades 10-12 on Forms Y, Z, and X. All the 375 items in the base item pool were concurrent calibrated to ensure the item parameters used in the data simulation were on the same scale. BILOG-MG 3 (Zimowski et al., 1996) was used to execute the item parameter estimation and the ability estimates were obtained via maximum-likelihood estimator (MLE) assuming the 3PLM. Once the level tests were constructed, the corresponding item parameter estimates of the selected items were used as the true item parameter estimates to simulate data for the level tests across Forms Y, Z, and X.

Next, the item response data were generated using a 3PLM for the nine grade level tests (three grades across the three forms). The probability of a correct response was calculated for each combination of examinees and items. If a random number generated from the standard uniform distribution was smaller than the probability of a correct response, the corresponding item response variable took on a value of 1. Otherwise the item response variable had a value of 0. Following this data generation scheme, the data were simulated for the nine level tests under each of the six conditions of proficiency distribution combinations.

To obtain the item parameter estimates under each of the six conditions, the nine level tests were separately calibrated using BILOG-MG 3 (Zimowski et al., 1996) assuming a 3PLM. The default priors of BILOG-MG 3 were used whenever applicable. Appendix B provides selected code used to conduct the separate calibrations for the level tests in BILOG-MG 3. In these calibrations, most of the χ^2 statistics for the model fit were not significant, showing evidence of reasonably good model fit for the IRT 3PLM. The separately calibrated proficiency estimates for the level tests were each scaled to have a mean of 0 and SD of 1 by the default setting of BILOG-MG 3. To develop a common proficiency scale for the nine tests across the three forms, some linking was needed. First, a vertical scale was created to put Grade 10 through Grade 12 on the same scale within form. Then, for the purpose of scale maintenance across grades and across forms, horizontal linking and vertical linking were conducted to put Form X onto the same scale with the baseline Form Y via interim Form Z. The next section describes the process of

conducting vertical scaling, followed by the scale maintenance processes of horizontal linking and vertical linking. The raw proficiency estimates of means and standard deviations prior to vertical scaling are reported in Table 4.2.

Constructing a Vertical Scale for Base Form Y

Vertical scales were constructed for the baseline Form Y under each of the six conditions of proficiency distribution combinations. The IRT scaling variability at this stage involved three sets of patterns of within-grade proficiency variability (constant, decreasing, and increasing). Thus three sets of vertical scales were established on the baseline Form Y for scale maintenance investigation: one set each for Conditions 1 and 2, Conditions 3 and 4, and Conditions 5 and 6. As described in Chapter III, three test forms were developed, each consisting of three level tests for Grades 10 to 12. Each level test consisted of item blocks that are unique for that grade level, vertical common items to be used for constructing a vertical scale or vertical linking, and horizontal items to be used for horizontal linking (see Table 3.1).

To construct a vertical scale for the baseline Form Y under the common-item test design, the level tests were developed in such a way that vertical common items between two adjacent grade levels helped establish the link across grades. After the item parameters and examinee proficiencies were estimated from separate calibration, the Stocking and Lord (1983) method was used to obtain the slopes and intercepts for the linear transformation to place all levels onto the same scale. The Stocking and Lord (1983) method uses the estimated QDs of the two grade levels and parameter estimates

for the vertical common items to obtain the slopes and intercepts under a test characteristic curve approach. The computer program ST (Zeng and Hanson, 1995) was used for obtaining linking constants. The resulting slopes and intercepts linking two adjacent levels for Forms Y, Z, and X under the six conditions of proficiency distribution are reported in Table 4.3. The slopes and intercepts for Form Y were used to create the baseline vertical scale. The slopes and intercepts for Forms Z and X were used to establish vertical scales for the corresponding forms in the process of scale maintenance via the vertical linking approach. As can be observed from the table, there are three sets of slopes and intercepts for the base Form Y, based on the three within-grade proficiency variability patterns. One set was for Conditions 1 and 2 with decreasing within-grade variability, one for Conditions 3 and 4 with constant within-grade variability, and the one for Conditions 5 and 6 with increasing within-grade variability. The values of the slopes and intercepts were decreasing as grade increased on Form Y under the six conditions. Similar trends on Forms Z and X can be observed, but the magnitudes of the slopes and intercepts on Forms Z and X were lower than those on Form Y.

On the baseline Form Y, Grade 10 was treated as the base grade and the other two grade levels were linked to the Grade 10 scale. Grade 12 needed multiple linear linking to the Grade 10 scale because estimated intercepts and slopes only existed for the adjacent levels. Therefore, to place Grade 12 on the same scale as Grade 10, it was linked to Grade 11 first and then to Grade 10, involving a total of two transformations. For example, the

following equations were used to calculate the new scale mean and SD for Grade 12 under Condition 1 for the baseline Form Y:

$$\overline{X}_{G12new} = (\overline{X}_{G12old} \times 0.986 + 0.379) \times 1.073 + 0.743 = \overline{X}_{G12old} \times 1.058 + 1.150$$

$$SD_{G12new} = SD_{G12old} \times 0.986 \times 1.073 = SD_{G12old} \times 1.058$$

The slopes and intercepts for the above transformations were taken from Table 4.3. Through this linking chain, a vertical scale was established, spanning Grades 10 to 12. Thus three sets of vertical scales were established on the baseline Form Y, to be used as the baseline scale for the scale maintenance across multiple forms Y, Z, and X.

Scale maintenance

Once the baseline vertical scale was established on the base Form Y, the next step was to transform the scale on the new Form X to the established vertical scale on Form Y through the interim Form Z (X to Z to Y). Two scale maintenance approaches were used to conduct the linking across forms, horizontal approach and vertical approach.

Horizontal Approach

As described in Chapter III, vertical common items were needed on Form Y to establish the baseline vertical scale, but no vertical common items were needed on Forms Z and X under the horizontal approach. The two sets of horizontal common items for the same grade across forms were used to obtain corresponding transformation constants between Forms Y and Z and between Forms Z and X. Table 4.4 reports the linking functions applied to link scales on Form X to the base scales on Form Y for the horizontal approach. The detailed process followed for the equating chain under the horizontal

approach was described in Chapter III. Through the equating chain, the parameter estimates for Form X were linked on the vertically adjusted scale on Form Y within a given grade. For example, to link the scale at Grade 12 on Form X to Form Y, the scale on Form X Grade 12 was first linked to the scale of Form Z using the linking functions in Table 4.4 obtained by the Stocking-Lord method on the horizontal common items between Form Z and Form X within Grade 12; next these results were transformed to the scales at Grade 12 on Form Y using the linking functions on the horizontal common items between Form Z and Form Y; then the vertical linking functions of Form Y through Grade 10 to Grade12 in Table 4.3 were used to place the scale onto the Grade 10 base scale on Form Y. Thus a total of four transformations were involved to place the scale at Grade 12 on Form X onto the base scale at Grade 10 on Form Y: Form X_G12 to Form Z_G12 to Form Y_G12 to Form Y_G11 to Form Y_G10. The following equations were used to calculate the new scale mean and SD for Grade 12 on Form X under Condition 1 via the horizontal approach:

$$\overline{X}_{G12new} = (((\overline{X}_{G12old} \times 0.964 + 0.312) \times 0.938 + 0.184) \times 0.986 + 0.379) \times 1.073 + 0.743$$

$$SD_{G12new} = SD_{G12old} \times 0.964 \times 0.938 \times 0.986 \times 1.073$$

To place Form X results onto the Form Y scale, the corresponding linking functions in Tables 4.3 and 4.4 were applied to proficiency estimates for each of the grade levels under each of the six conditions. After linking, the proficiency parameter estimates were on the Form Y vertical scale. No further adjustment was needed.

Vertical Approach

The vertical approach involved establishing vertical scales on each of the three forms and linking the vertical scales. Under this maintenance approach in this dissertation, three separate vertical scales were constructed, one for Form Y, one for Form Z, and the other for Form X. Next the vertical scale developed for Form X was linked to the vertical scale on Form Y.

To achieve this linking across the three vertical scales from Form X to Form Y via Form Z, common items were used to identify the linking relationship. These were the same sets of horizontal common items used in the horizontal approach; under the vertical approach, these items were used to link vertical scales, instead of linking tests within the same grade. As described in Chapter III, the linear transformation was estimated using the Stocking-Lord method with all the common items combined across all grade levels, after placing all items onto the vertical scale. Table 4.4 reports the linking functions applied to vertical linking approach. The linear function is labeled as “all grades” in Table 4.5. For example, to link the scale at Grade 12 on Form X to Form Y under the vertical approach, the scale of Grade 12 on Form X was first linked to the scale of Grade 10 on the same form using the linking functions on vertical common items through Grades 10 to 12 on Form X in Table 4.3. Thus a vertical scale was established on Form X through Grades10 to 12. Next the results were transformed to the scales on Form Z using the all-grade linking functions on all the horizontal common items through Grades 10 to 12 between Form Z and Form X in Table 4.5. Then the all-grade linking functions between

Form Z and Form Y through Grades 10 to 12 in Table 4.5 were used to place the scale onto the Grade 10 base scale on Form Y. Thus a total of four transformations were involved to place the scale at Grade 12 on Form X onto the base scale at Grade 10 on Form Y: Form X_G12 to Form X_G11 to Form X_G10 to Form Z to Form Y_G10. Comparing with the four transformations involved in the horizontal approach to place the scale of Grade 12 on Form X to the scale of Grade 10 on Form Y, the linking chain under the vertical approach involved a total of four transformations with a different path. The following equations were used to calculate the new scale mean and SD for Grade 12 on Form X under Condition 1 via vertical approach:

$$\overline{X}_{G12new} = (((\overline{X}_{G12old} \times 0.964 + 0.312) \times 0.987 + 0.723) \times 0.943 + 0.274) \times 0.924 + 0.212$$

$$SD_{G12new} = SD_{G12old} \times 0.964 \times 0.987 \times 0.943 \times 0.924$$

To place Form X results onto the Form Y scale under the vertical approach, the corresponding linking functions in Tables 4.3 and 4.5 were applied to proficiency estimates for each of the grade levels under each of the six conditions.

After linking across the three forms via the horizontal approach or the vertical approach, the proficiency parameter estimates were on the Form Y vertical scale. These multiple linkings potentially may allow error to accumulate. Through these multistage linkings, all the levels were placed onto the same vertical scale for each of the six conditions. In the following sections, results are presented for each pair of the six conditions of proficiency distributions: Conditions 1 and 2, Conditions 3 and 4, and Conditions 5 and 6.

With regard to the first research question of this dissertation, the resulting Pearson product-moment coefficients, root-mean square error (RMSE), and mean absolute difference (MAD) were used to evaluate how adequate the two scale maintenance approaches are in recovering the true examinee proficiencies on the new Form X and the interim Form Z under each of the six conditions of proficiency distribution. With regard to the second research question (the effect of the two scale maintenance approaches on the resulting scales and growth interpretation) three properties were employed: grade-to-grade growth, grade-to-grade variability, and separation of grade distributions. The indices obtained via the two scale maintenance approaches on Forms X and Z (i.e., mean, standard deviation, mean difference between adjacent grades, and effect size) were reported, and comparisons were made under each pair of the six conditions. Finally, the ratios of RMSE and the ratios of MAD between Form X and Form Z within each of the two scale maintenance approaches were reported under each pair of conditions, as well as the ratios of RMSE and the ratios of MAD between the horizontal and vertical approaches within Form X and within Form Z. Comparisons among these ratios within forms and within scale maintenance approaches helped address the third research question (i.e., under which condition is one approach preferable to the other). In this simulation study, the values of all the reported indices were average values of the corresponding indices over 100 replications with sample sizes of 2000 per form. In all the figures in this chapter, the dashed lines represent the results on the interim Form Z; the solid lines represent the results on the new Form X.

Conditions 1 and 2

Conditions 1 and 2 were sets of proficiency distributions where the true vertical scale's within-grade variability remained constant as grade increased. The same baseline vertical scale on Form Y was developed under the two conditions. The difference between Conditions 1 and 2 exists in the examinee sample characteristics across the three forms. Condition 1 represented small group differences across Forms Y, Z, and X; Condition 2 represented large group differences across the three forms. Table 3.7 lists the means and SDs for the underlying proficiency distribution for each grade level under Conditions 1 and 2, with the true means increasing from Grades 10 to 12 and the within-grade variability remaining constant at 1 across grades. The examinee group differences between adjacent forms (Y and Z, Z and X) had a difference value of 0.25 under Condition 1 (small) and a difference value of 0.5 under Condition 2 (large).

Adequacy in Recovering the True Proficiency

Correlations

Table 4.6 lists the average means and standard deviations of correlations under Conditions 1 and 2 over 100 replications. Across Grades 10-12 the two scale maintenance approaches yielded very similar means and standard deviations of correlations within grade level within form under both conditions. In general the mean correlations ranged from 0.901 to 0.935, indicating that the proficiency scales developed by the horizontal approach and by the vertical approach matched the true proficiency scales pretty well. However, when single linking (on Form Z) or multiple linking (on Form X) across forms

on vertical scales was conducted, some slight differences could be observed between the horizontal approach and the vertical approach.

Figure 4.1 contrasts the average correlations via the two scale maintenance approaches across grades under Conditions 1 and 2. By comparing the results between the two scale maintenance approaches within form under the two conditions, the vertical approach yielded slightly higher mean correlations than the horizontal approach. The magnitude of the average correlations was larger on the interim Form Z with single linking to the base Form Y. When multiple linking was conducted on Form X, the magnitude of average correlations was slightly smaller under both conditions, indicating that more drift occurred with multiple linking across three forms.

The corresponding results under Condition 2 followed a similar trend as Condition 1, but the average correlations were smaller than those under Condition 1, as expected, since the examinee group differences across forms were larger under Condition 2. Furthermore, the decreasing trend of the average correlations across forms was more noticeable under Condition 2, indicating that larger examinee group difference across forms introduced more drifts along with the impact of multiple linking. For instance, under Condition 1, the average correlations across grades via the horizontal approach decreased from 0.930 on Form Z to 0.922 on Form X. Under Condition 2, the average of mean correlations across grades via the horizontal approach decreased from 0.924 on Form Z to 0.905 on Form X. Regardless of form and condition, the horizontal approach and the vertical approach yielded similar correlations.

RMSE and MAD

The means and standard deviations of RMSE via the two approaches under Conditions 1 and 2 over 100 replications are reported in Table 4.7.The average means and standard deviations of RMSE across Grades 10-12 are reported in Table 4.7. Within each of the two forms under both conditions, the horizontal approach yielded smaller RMSE than the vertical approach. As multiple linking increased across Forms X, Z and Y, the RMSEs of both horizontal and vertical approaches increased under the two conditions. For instance, the average RMSE on Form X via the horizontal approach increased to 0.401 from 0.375 on Form Z. The extent of increase from Form Z to Form X under Condition 1 (from 0.375 to 0.401 via the horizontal approach and from 0.390 to 0.414 via vertical approach), was smaller than that under Condition 2 (from 0.388 to 0.456 via the horizontal approach and from 0.393 to 0.477 via the vertical approach). Furthermore, by comparing the corresponding results between the two conditions, the magnitudes of RMSEs under Condition 2 were larger than those under Condition 1. A multiple linking increased on Form X, the average RMSE increased under Condition 2 from 0.388 to 0.456 via the horizontal approach and from 0.393 to 0.477 via the vertical approach. These findings indicated that when the examinee group difference increased under Condition 2, more bias and random errors were introduced on both Form Z and Form X, especially on Form X.

Figure 4.2 visually demonstrates RMSE changes between the horizontal approach and the vertical approach as grade increased under Conditions 1 and 2. On the interim

Form Z under Condition 1 as grade increased the RMSE via the horizontal approach decreased at Grade 11 but remained flat as grade increased to Grade 12; the RMSE via the vertical approach decreased at Grade 11 but increased at Grade 12. The horizontal approach and the vertical approach yielded pretty similar RMSEs at Grades 10 and 11 but the difference between the two approaches was relatively larger at Grade 12. This indicates that the horizontal approach introduced less bias or random error than the vertical approach as grade increased, especially at Grade 12 on the interim Form Z with single linking across Forms Z and Y. With multiple linking on Form X under Condition 1, it was observed that the magnitude of RMSEs is larger than that on Form Z. As grade increased, the RMSE via the horizontal approach decreased at Grade 11 and then increased slightly at Grade 12; the RMSE via the vertical approach increased at Grade 11 and then decreased at Grade 12 with similar values as Grade 10. On Form X under Condition 1, the performances of the horizontal and the vertical approaches were pretty consistent at Grades 10 and 12 while fluctuation existed at Grade 11 showing a better performance on the horizontal approach.

As shown in Figure 4.2, on the interim Form Z under Condition 2, as grade increased, the performance of the horizontal approach was pretty consistent across Grades 10 to 12; the performance of the vertical approach was very similar to the horizontal approach at Grades 10 and 11 with slight increase at Grade 12. In general, the horizontal approach and the vertical approach displayed similar performances at Grades 10 and 11 while the horizontal approach performed better at Grade 12, indicating that the

vertical approach might introduce more bias at higher grades. Since the baseline grade is Grade 10, more error was expected at Grade 12. As compared to the results under Condition 1, as grade increased, the discrepancy at Grade 12 between the two scale maintenance approaches under Condition 2 was slightly smaller but the magnitude of the RMSEs was larger than under Condition 1. As can be seen in Figure 4.2, on the new Form X under Condition 2, the curves of both the horizontal and the vertical approaches increased systematically as grade increased, and the magnitudes of RMSEs are much larger than those on Form Z. This indicates that more bias was introduced by multiple linking on the vertical scale across the three forms and by larger examinee group differences across the three forms. In general, the horizontal approach performed better than the vertical approach through Grades 10 to 12.

The means and standard deviations of MAD via the two approaches under Conditions 1 and 2 over 100 replications are reported in Table 4.8 and are plotted in Figure 4.3. By comparing with the results of RMSEs, as can be seen in Figures 4.2 and 4.3, the overall magnitudes of MADs were smaller than those of RMSEs under both conditions via corresponding scale maintenance approaches for each of the two forms. As expected, similar trends of MAD changes as those of the corresponding RMSE changes were observed as grade increased within form under each of the two conditions. However, the extent of increase and decrease were relatively smaller in MAD than in RMSE. For instance, the MAD difference between the horizontal approach and the vertical approach at Grade 11 on Form X under Condition 1 was around 0.021 (0.306

versus 0.327); the corresponding RMSE difference between the two approaches was around 0.030 (0.394 versus 0.424). In addition, the magnitude of MAD differences between the interim Form Z and the new Form X under Condition 2 was smaller than that of RMSE differences between the two forms. For instance, the MAD differences at Grade 11 between Forms Z and X were 0.058 via the horizontal approach (0.306 versus 0.364) and 0.048 via the vertical approach (0.327 versus 0.375); the RMSE differences at Grade 11 between the two forms were 0.072 via the horizontal approach (0.394 versus 0.466) and 0.056 via the vertical approach (0.424 to 0.480).

Figure 4.4 plots the average RMSEs and MADs under Conditions 1 and 2. The RMSEs and the corresponding MADs under the two conditions yielded almost parallel results, and the RMSEs were always larger than the corresponding MADs, especially under Condition 2. The RMSEs and MADs were larger under Condition 2 than under Condition 1, indicating that large group difference across forms introduced more error in the performance of the two scale maintenance approaches. In addition, the RMSEs or MADs on the interim Form Z via the two scale maintenance approaches were very close under the two conditions. The extent of increase is more noticeable on the new Form X under Condition 2, indicating that the greater the group differences across forms the larger the errors when multiple linking increased across three forms in this study.

Growth Statistics

As mentioned in Chapter III, three indices were used to describe the growth trends established by the two scale maintenance approaches in linking across forms. They are

grade-to-grade growth, grade-to-grade variability, and separation of grade distributions. The consistency of the estimated proficiency scales with the true proficiency distribution patterns of these three properties were compared for the two approaches.

Table 4.9 and Table 4.10 report the grade-to-grade means, standard deviations, mean differences and effect sizes for the proficiency scales via the two approaches averaged under Conditions 1 and 2 over 100 replications. Figure 4.5 through Figure 4.7 provide the results of proficiency estimates via the two scale maintenance approaches across forms under Conditions 1 and 2.

Grade-to-grade growth

The pattern of grade-to-grade growth is represented by means and mean differences between adjacent grades within form. In estimating means regardless of conditions and forms, as Tables 4.9 and 4.10 indicate, both of the scale maintenance approaches produced an increasing trend from Grade 10 to Grade 12, showing growth from lower to higher levels. However, the results between the two maintenance approaches provided somewhat different results on Forms Z and X under each of the two conditions. To examine the differences more closely, Figure 4.5 was constructed to illustrate the differences on the mean estimates on Forms Z and X under each of the two conditions. True_Form Z represents the true mean on Form Z; Horizontal_Form Z represents the mean estimates via the horizontal approach on Form Z; Vertical_Form Z represents the mean estimates via the vertical approach on Form Z. True_Form X represents the true mean on Form X; Horizontal_Form X represents the mean estimates

via the horizontal approach on Form X; Vertical_Form X represents the mean estimates via the vertical approach on Form X.

Under Condition 1, as can be observed from Figure 4.5, on both Forms Z and X, the horizontal approach appeared to overestimate the true grade means while the vertical approach appeared to underestimate the true grade means; the extent of overestimation and underestimation tended to be larger as grade increased, especially on Form X. A possible explanation for the larger extent of over- and under-estimations via the two approaches on Form X would be that multiple linking caused more estimation discrepancies on Form X, especially at the higher grade with more transformations involved through a linking chain. For instance, at Grade 12, the difference between the true mean (1.272) and mean estimate via the horizontal approach (1.345) on Form Z was 0.073, which involved three transformations (G12_Form Z to G12_Form Y to G11_Form Y to G10_Form Y); the difference between the true mean (1.523) and the mean estimate via the horizontal approach (1.656) on Form X increased to 0.133, which involved one more transformation (G12_Form X to G12_Form Z to G12_Form Y to G11_Form Y to G10_Form Y). In terms of the two maintenance approaches, both approaches were able to capture the increasing trend in performance, in accordance with larger magnitude of the true means on Form X. With regard to the value discrepancy between the true means and the mean estimates, the horizontal approach tended to perform better than the vertical approach. If preference for the maintenance of a vertical scale is to be able to capture similar characteristics on the new Form X as on the true growth trends, then the vertical

approach appeared to produce similar growth trends on the new Form X in terms of relative positions of each grade compared to the true positions.

Under Condition 2, as showed in Table 4.10, both approaches tended to underestimate the true mean at each grade on each form, except the horizontal approach at Grade 10 on Forms Z and X. As depicted in Figure 4.5, the extent of underestimation was much more evident on the new Form X; the mean estimates on the interim Form Z appeared very similar between the true means and the mean estimates via the two approaches. By comparing the results with those under Condition 1, the two approaches produced better mean estimates on the interim Form Z under Condition 2. However, when multiple linking and larger examinee group differences were involved between the baseline Form Y and the new Form X, larger discrepancies were observed between the true means and the mean estimates at each grade on Form X. Regardless of the mean estimate at Grade 10 via the horizontal approach on Form X, it is hard to tell which approach was doing better because the mean estimates via the two approaches were very close. However, with regard to how well similar characteristics of the true growth trend on the new Form X were captured, the vertical approach showed better performance in terms of relative positions of each grade as compared to the true ones. From this perspective, the vertical approach captured more similar characteristics of the true growth trend under both conditions, but distances between the true growth trend and the estimated growth trend were larger under Condition 2, due primarily to larger examinee group differences across forms.

Grade-to-grade Variability

The standard deviation estimates via the two approaches on each of the two forms under Conditions 1 and 2 are reported in Tables 4.9 and 4.10. Under both conditions, the within-grade variability of the true vertical scale remained constant as grade increased on each of the three forms. With respect to the within-grade variability, underestimation trends can be detected via both approaches across forms under the two conditions, as shown in Tables 4.9 and 4.10. Figure 4.6 provides visual observations about the extent of underestimation of the true standard deviations via the two approaches under Conditions 1 and 2.

Under Condition 1, the standard deviation estimates via the horizontal approach tended to fluctuate as grade increased, decreasing at Grade 11 and then increasing at Grade 12; the standard deviation estimates via the vertical approach showed a slight decrease from lower to higher grades. It can be observed that the resulting SD estimates via the horizontal approach for the interim Form Z and the new Form X were very close; so were the resulting SD estimates via the vertical approach. This suggests that examinee group difference across forms had almost no impact on the resulting estimates of standard deviations via the two approaches. By comparing the resulting estimates between the horizontal and the vertical approaches for Form Z and Form X, both approaches produced similar estimates at Grades 10 and 11; as grade increased to Grade 12, the resulting estimates via the vertical approach tended to deviate more from the true value and those via the horizontal approach showed much smaller distance from the true value. The

vertical approach yielded a closer approximation compared to the true constant within-grade variability pattern, but the distances are larger than the horizontal approach on both Form Z and Form X.

The standard deviation estimates via the two approaches across forms under Condition 2 provided similar results to those under Condition 1. In Figure 4.6 slight differences between the two approaches can be detected. Via the vertical approach, the resulting estimates at Grades 10 and 11 were closer to the true values under Condition 2 than under Condition 1. Via the horizontal approach, relatively larger distances were observed between Form Z and Form X under Condition 2, and the resulting estimates across grades tended to be flatter under Condition 2. Under Condition 2, it is hard to decide which scale maintenance approach produced a within-grade variability pattern that is more similar to the true constant within-grade variability pattern across grades. But the horizontal approach tended to yield flatter trends on both the interim Form Z and the new Form X.

Separation of Grade Distributions

The separation of grade distributions can be observed through the effect size estimates between adjacent grade levels. The effect size estimates via the two approaches on each of the two forms under Conditions 1 and 2 are reported in the last section of Tables 4.9 and 4.10.

Under Condition 1, as reported in Table 4.9, the true effect sizes showed a decreasing trend as grade increased, indicating decelerated growth from lower to higher

grades. Both approaches captured the decreasing pattern of the true effect size across forms. However, both approaches tended to overestimate the true effect size at each grade on each form, except the vertical approach at Grades11/12 on Form X. On the interim Form Z, the two approaches yielded almost parallel trends to the true effect size pattern; the extent of overestimation via the vertical approach is smaller than the horizontal approach and thus showed a better performance than the horizontal approach. On the new Form X when multiple linking was involved, more fluctuation can be detected in Figure 4.7. The decreasing pattern was much flatter via the horizontal approach and larger discrepancies can be observed between the true effect size and the estimate at Grades11/12. The vertical approach showed an underestimate at Grades11/12 on Form X, but the distance from the true effect size is much smaller than that via the horizontal approach. In general, under Condition 1, the vertical approach resulted in a more similar trend to the true effect size pattern on both the interim Form Z and the new Form X; the horizontal approach showed larger discrepancies from the true effect size pattern, especially at higher Grades 11/12 on the new Form X.

Under Condition 2, more fluctuations were observed on Form X in Figure 4.7. As reported in Table 4.10, both approaches tended to overestimate the true effect size on the interim Form Z. Figure 4.7 shows that the vertical approach yielded an almost parallel pattern to the true effect size pattern on Form Z; more deviation can be observed via the horizontal approach on Form Z. With regard to the new Form X, the resulting estimate of Grades 10/11 via the vertical approach was pretty close to the true effect size, and an

overestimate can be observed for Grades 11/12. More fluctuations via the horizontal approach on the new Form X were observed in Figure 4.7. The horizontal approach yielded an almost flat pattern of effect size, with a large discrepancy from the true effect size at Grades 10/11. In general, on the interim Form Z involving single linking, the two scale maintenance approaches yielded decreasing patterns with slight overestimation of the true pattern, and the vertical approach tended to show a more similar pattern to the true one. On the new Form X with multiple linkings and larger examinee group differences, the vertical approach showed better performance than the horizontal approach across grades, especially at Grades 10/11.

Ratios of RMSE and MAD

The results via the two scale maintenance approaches in the above two sections are mixed under Conditions 1 and 2. Discrepancies were detected via each approach across forms under each condition, indicating that multiple linking across three forms had an impact on the resulting indices: correlation, RMSE, MAD, mean, standard deviation, and effect size. Furthermore, in comparisons of the results via the two approaches between the two conditions, examinee group differences did show some impact on the resulting indices, especially under Condition 2 with larger group differences across forms. It is hard to decide whether the discrepancies in the resulting indices between the horizontal approach and the vertical approach were due to different approaches only or due to larger examinee group difference across forms only or due to both approach difference and group difference. In order to further investigate under which condition one

approach is preferable to the other, the ratios of the average RMSEs and the ratios of the average MADs between Form Z and Form X via the two scale maintenance approaches are plotted in Figure 4.8, as well as the ratios of the average RMSEs and the ratios of the average MADs between the two approaches for each Form Z and Form X. V/H_Form Z represents the ratio of the average RMSEs or MADs between the vertical approach and the horizontal approach on Form Z; V/H_Form X represents the ratio of the average RMSEs or MADs between the vertical approach and the horizontal approach on Form X. X/Z_Horizontal represents the ratio of the average RMSEs or MADs between Form Z and Form X via the horizontal approach; X/Z_Vertical represents the ratio of the average RMSEs or MADs between Form X and Form Z via the vertical approach. The average RMSEs and MADs are reported in Tables 4.7 and 4.8. As can be observed in Figure 4.8, under both conditions, the values of the RMSE ratios and the MAD ratios were very consistent for each of the four sets.

Under Condition 1, in comparisons of the Vertical/Horizontal ratios between the two forms, the ratios of V/H_Form Z was slightly higher than that of V/H_Form X, indicating that the vertical approach tended to introduce less error than the horizontal approach when multiple linking was involved on the new Form X. By comparing the X/Z ratios between the two scale maintenance approaches, the ratio of X/Z_Horizontal were higher than that of X/Z_Vertical, which confirmed that the extent of error increase via the vertical approach was less than that via the horizontal approach when multiple linking was involved on Form X. In addition, the magnitudes of the ratios of

X/Z_Horizontal and X/Z_Vertical were larger than those of V/H_Form Z and V/H_Form X. This suggests that multiple linking introduced more error in the process of maintaining vertical scales across the three forms and had more impact on the performance of the horizontal approach than on the performance of the vertical approach. For instance, the RMSE on Form X was 1.117 times larger than that on Form Z via the horizontal approach; the RMSE on Form X was 1.062 times larger than that on Form Z via the vertical approach.

Under Condition 2, in comparisons of the Vertical/Horizontal ratios between the two forms, the ratios of V/H_Form Z was slightly smaller than that of V/H_Form X, indicating that the vertical approach tended to introduce a bit more error than the horizontal approach when larger examinee group difference existed between Form Z and Form X. By comparing the X/Z ratios between the two scale maintenance approaches, the ratio of X/Z_Horizontal were smaller than that of X/Z_Vertical, which confirmed that the extent of error increase via the horizontal approach was less than that via the vertical approach when examinee group difference became larger between Form Z and Form X under Condition 2. In addition, the magnitudes of the ratios of X/Z_Horizontal and X/Z_Vertical were larger than those of V/H_Form Z and V/H_Form X. This suggests that multiple linking introduced more errors in the process of maintaining vertical scales across the three forms as under Condition 1 and larger examinee group difference between Forms Z and X had more impact on the performance of the vertical approach than on the performance of the horizontal approach. For instance, the RMSE on Form X

was 1.177 times larger than that on Form Z via the horizontal approach; the RMSE on Form X was 1.212 times larger than that on Form Z via the vertical approach.

By comparing the ratios between Conditions 1 and 2 in Figure 4.8, the discrepancies of X/Z_Horizontal and X/Z_Vertical between the two conditions were more noticeable than those of V/H_Form Z and V/H_Form Z. This indicates that multiple linking across the three forms did have an impact on the resulting proficiency estimates conversions by the scale maintenance approaches but the extent was less than the impact of large examinee group difference on the resulting proficiency estimates conversions by the two approaches.

Conditions 3 and 4

Under Conditions 3 and 4, the true means increased from one grade to the next, in the same manner as Conditions 1 and 2, but the within-grade variability decreased from an SD of 1 for Grade 10 to an SD of .631 for Grade 12, a 40% drop across the three grades. The same baseline vertical scale on Form Y was developed under the two conditions. As with Conditions 1 and 2, the difference between Conditions 3 and 4 existed in examinee sample characteristics across the three forms. Condition 3 represented small group difference across Forms Y, Z, and X; Condition 4 represented large group difference across the three forms.

Adequacy in Recovering the True Proficiency

Correlations

Table 4.11 lists the means and standard deviations of correlations under Conditions 3 and 4 over 100 replications. The means and standard deviations of correlations across Grades 10-12 were averaged and labeled as “Average” in Table 4.11. Similar to what was observed under Conditions 1 and 2, the two scale maintenance approaches yielded very similar means and standard deviations of correlations within level test on each form under both conditions. The magnitude of the mean correlations under Conditions 3 and 4 was slightly smaller than that under Conditions 1 and 2. In general the mean correlations ranged from 0.900 to 0.927, indicating that the proficiency scales developed by the horizontal approach and by the vertical approach matched the true proficiency scales pretty well. However, when single linking (on Form Z) or multiple linking (on Form X) across forms on vertical scales was conducted, some slight differences were observed between the horizontal approach and the vertical approach.

Figure 4.9 plots the average of mean correlations across grades under the two conditions. By comparing the results between the two scale maintenance approaches on each form under the two conditions with the decreasing within-grade variability pattern, the vertical approach again yielded slightly higher average correlations than the horizontal approach. The magnitude of the average correlations was larger on the interim Form Z when linking the interim Form Z to the baseline Form Y. When multiple linking was conducted on Form X, the magnitude of average correlations was slightly smaller

under both conditions, indicating that more drift occurred with multiple linking across the three forms. Furthermore, the extent of decrease was more noticeable under Condition 4, indicating that larger examinee group differences across the three forms introduced more errors along with the impact of multiple linking. For instance, under Condition 3, the average correlations via the horizontal approach decreased from 0.921 on Form Z to 0.915 on Form X; under Condition 4, the average correlations via the horizontal approach decreased from 0.917 on Form Z to 0.906 on Form X.

RMSE and MAD

The means and standard deviations of RMSEs and MADs via the two approaches under Conditions 3 and 4 over 100 replications were reported in Tables 4.12 and 4.13. The means and standard deviations of RMSE across Grades 10-12 were averaged and labeled as “Average” in the tables.

Figure 4.10 visually demonstrates RMSE trends between the horizontal approach and the vertical approach as grade increases under Conditions 3 and 4. Regardless of approaches and forms, decreasing patterns can be observed as grade increased under both conditions, which is different from the fluctuating patterns from low to high grades under Conditions 1 and 2. Under Condition 3, as grade increased, the horizontal approach and the vertical approach yielded similar decreasing patterns of RMSEs on Forms Z and X but the difference between the two approaches was relatively larger on Form X. The horizontal approach tended to introduce less bias or random error than the vertical approach, especially when multiple linking was involved on Form X.

As seen in Figure 4.10, under Condition 4, the performance of the horizontal approach and the vertical approach were consistent, showing a gradually decreasing pattern as grade increased on the interim Form Z. On Form X with multiple linking, the decreasing pattern via the horizontal approach became flatter than that via the vertical approach. The vertical approach yielded an almost parallel decreasing pattern as that on Form Z, but the magnitude of RMSEs was larger on Form X. In general, the horizontal approach and the vertical approach displayed similar performances with single linking on the interim Form Z; on Form X with multiple linking across the three forms, the horizontal approach performed better at Grade 10 and the vertical approach showed better performance at Grade 12.

The means and standard deviations of MADs via the two approaches under Conditions 3 and 4 over 100 replications are reported in Table 4.13 and are plotted in Figure 4.11. By comparing with the results of RMSEs, as can be seen in Figures 4.10 and 4.11, the overall magnitudes of MADs were smaller than those of RMSEs under both conditions via the two corresponding scale maintenance approaches on each of the two forms. As expected, MAD trends were similar to those of corresponding RMSE trends as grade increased on each form under each of the two conditions.

Figure 4.12 plots the average of mean RMSEs and MADs across grades under the two conditions. Similar to Conditions 1 and 2, the magnitude of MADs was smaller than that of RMSEs, but both indices showed almost parallel results across conditions. As depicted in Figure 4.12, on each of the two forms under both conditions, the horizontal

approach yielded smaller RMSE and MAD than the vertical approach. When multiple linking was involved across Forms X, Z and Y, the RMSEs and the MADs of both the horizontal and vertical approaches increased under the two conditions. The extent of increase on Form X via the vertical approach was more noticeable.

Growth Statistics

Tables 4.14 and 4.15 report the grade-to-grade means, standard deviations, mean differences and effect sizes for the proficiency scales via the two approaches averaged under Conditions 3 and 4 over 100 replications. Figure 4.13 through Figure 4.15 are constructed for visual observation of the results of proficiency estimates via the two scale maintenance approaches across forms under Conditions 3 and 4 (mean estimates, standard deviation estimates, and effect size estimates).

Grade-to-grade growth

The pattern of grade-to-grade growth is represented by means and mean differences between adjacent grades within form. Similar to what was observed with Conditions 1 and 2, the two scale maintenance approaches under investigation were able to capture the increasing trend in means across grades under Conditions 3 and 4, regardless of single or multiple linking. This indicates higher performance at higher grades. Yet the results between the two maintenance approaches are somewhat different on Forms Z and X under each of the two conditions. To examine the differences more closely, Figure 4.13 was constructed to illustrate the differences in the mean estimates on

Forms Z and X under each of the two conditions, in the same layout as the figures created for Conditions 1 and 2.

In Figure 4.13, under Condition 3, on both Forms Z and X, the horizontal approach showed very close mean estimates to the true grade means while the vertical approach appeared to overestimate the true grade means. The extent of overestimation tended to be consistent as grade increased on the interim Form Z as well as the new Form X, with somewhat more evident overestimation on Form X when multiple linking was involved. Both approaches were able to capture the increasing trend of the true performance, with larger magnitude of true means on Form X as designed. The horizontal approach tended to perform better than the vertical approach in recovering the true means. The horizontal approach appeared to produce very close growth trends on the new Form X in terms of relative positions of each grade.

Similar to what can be observed on Condition 3 in Figure 4.13, the horizontal approach showed better performance on both forms under Condition 4, except at Grade 12 on the new Form X. The vertical approach appeared to overestimate the true means on both forms, and the extent of overestimation tended to be less as grade increased. The vertical approach showed in general more consistent performance with regard to how well it captured similar characteristics of the true growth trend in terms of relative positions of each grade on Form Z and Form X.

Grade-to-grade Variability

The standard deviation estimates via the two approaches on each of the two forms under Conditions 3 and 4 are reported in Tables 4.14 and 4.15. Under both conditions, the within-grade variability of the true vertical scale decreased as grade increased on each of the three forms. With respect to the within-grade variability, a tendency to underestimate can be detected via both approaches across forms under the two conditions, which is similar to what was observed under Conditions 1 and 2. Figure 4.14 provides visual observations about the extent of underestimation of the true standard deviations via the two approaches under Conditions 3 and 4.

Under Condition 3, the horizontal approach and the vertical approach showed similar performance at Grades 10 and 11on the interim Form Z. At Grade 12, the SD estimate via the vertical approach tended to deviate more from the true SD while the SD estimate via the horizontal approach tended to be close to the true SD. On the new Form X, the two scale maintenance approaches performed similarly at Grades 10 and 11; but the underestimation was more evident than on the interim Form Z. Discrepancies between the resulting SD estimates via the two approaches existed at Grade 12 on Form X but were less evident than that on the interim Form Z. The resulting estimates via the horizontal approach for the interim Form Z and the new Form X were almost parallel with larger magnitude on Form Z. The SD estimates via the vertical approach on the two forms were almost parallel at Grades 10 and 11, but the decreasing trend on Form X is less than that on Form Z. This suggests that multiple linking across the three forms may

have more impact on the resulting estimates of standard deviations via the vertical approach but less on the results via the horizontal approach, when the within-grade variability has a decreasing pattern as grade increases.

The standard deviation estimates via the two approaches across forms under Condition 4 provided similar results at Grades 10 and 11 on the interim Form Z as those under Condition 3, but the resulting SD estimates via the vertical approach deviated more from the true SD at Grade 12 on Form Z. As can be observed in Figure 4.14, the horizontal approach and the vertical approach yielded similar results at Grades 10 and 12 on the new Form X; discrepancies between the two approaches were more noticeable at Grade 11 on Form X. The pattern of SD estimates across grades on Form X via the vertical approach tended to be flat at Grades 10 and 11, but the extent of decrease at Grade 12 was larger than that via the horizontal approach. Via the horizontal approach, relatively larger distances were observed between Form Z and Form X under Condition 4.

It is hard to decide which scale maintenance approach produced more accurate within-grade variability patterns for the true decreasing pattern across grades under the two conditions. As depicted in Figure 4.14, the horizontal approach showed more consistent or parallel decreasing trends across grades on the two forms under Conditions 3 and 4; more fluctuations across grades can be observed in the resulting SD estimates via the vertical approach, for instance, at Grade 12 on Form X under Condition 3.

Separation of Grade Distributions

The separation of grade distributions can be observed through the effect size estimates between adjacent grade levels. The effect size estimates via the two approaches on each of the two forms under Conditions 3 and 4 are reported in the last section of Tables 4.14 and 4.15.

Under Condition 3, as reported in Table 4.14, the true effect sizes showed a decreasing trend as grade increased, indicating decelerated growth from lower to higher grades. Both approaches captured the decreasing pattern of effect sizes across forms. Similar to what was observed under Conditions 1 and 2, both approaches tended to overestimate the true effect size at each grade on each form, except the vertical approach at Grades11/12 on Form X. On the interim Form Z, the vertical approach yielded an almost parallel trend to the true effect size pattern. The extent of overestimation via the horizontal approach was smaller than the vertical approach and thus showed better estimates than the vertical approach. On the new Form X when multiple linking was involved, more fluctuations were detected in Figure 4.15. The decreasing pattern was much flatter via the horizontal approach, and large discrepancies were observed between the true effect size and the estimate at Grades 11/12. Both approaches showed overestimates at Grades 10/11, and the overestimation of the vertical approach was more noticeable than the horizontal approach on Form X. But the vertical approach yielded results very close to the true effect size at Grade 12 and thus resulted in a larger decreasing pattern of effect sizes than the true decreasing pattern as grade increased. In

general, under Condition 3 the horizontal approach performed better than the vertical approach on the interim Form Z but fluctuations were more evident on Form X. Thus it is hard to decide which scale maintenance approach performed better on Form X.

Under Condition 4, the two scale maintenance approaches showed overestimation of the true effect size as grade increased on the interim Form Z with single linking. As depicted in Figure 4.15, the horizontal approach yielded an almost parallel pattern as the true effect size pattern on Form Z; the vertical approach yielded more accurate estimates of the true effect sizes. With regard to the new Form X involving multiple linkings, the vertical approach underestimated the true effect sizes; the horizontal approach had overestimated results at Grade 10/11 and underestimated results at Grade 11/12. In general, the vertical approach on the new Form X tended to show a more similar pattern to the true one. Similar to Condition 3, multiple linkings across forms tended to have more impact on the horizontal approach under Condition 4.

Ratios of RMSE and MAD

The results via the two scale maintenance approaches in the above two sections are mixed under Conditions 3 and 4. Discrepancies were detected via each approach across forms under each condition, indicating that multiple linkings across the three forms had an impact on the resulting indices: correlation, RMSE, MAD, mean, standard deviation, and effect size. It is hard to decide whether the discrepancies in the resulting indices between the horizontal approach and the vertical approach were due to different approaches only, to larger examinee group differences across forms only, or to both the

approach difference and group difference. In this section, to further investigate under which condition one approach is preferable to the other, Figure 4.16 provides visual observations of the ratios of the average RMSEs and the ratios of the average MADs in the same layout as Figure 4.8 under Conditions 1 and 2. The average RMSEs and MADs are reported in Tables 4.12 and 4.13. As can be observed in Figure 4.16, under both conditions, the values of the RMSE ratios and the MAD ratios were very consistent for each of the four sets. All the ratios were greater than 1, indicating that the vertical approach introduced more error than the horizontal approach and multiple linkings on Form X introduced more errors than single linking on Form Z.

Under Condition 3, by comparing the Vertical/Horizontal ratios between the two forms, the ratios of V/H_Form X was slightly higher than that of V/H_Form Z, indicating that the vertical approach tended to introduce a bit more error than the horizontal approach when multiple linkings were involved on the new Form X. By comparing the X/Z ratios between the two scale maintenance approaches, the ratio of X/Z_Horizontal were lower than that of X/Z_Vertical, which confirmed that the error increase via the vertical approach was a bit more than that via the horizontal approach when multiple linkings were involved on Form X. In addition, it can be observed that the magnitudes of the ratios of X/Z_Horizontal and X/Z_Vertical were slightly larger than those of V/H_Form Z and V/H_Form X. However, these differences were so small that it is hard to tell which scale maintenance approach is better under Condition 3.

Under Condition 4, by comparing the Vertical/Horizontal ratios between the two forms, the ratios of V/H_Form Z and the ratios of V/H_Form X were very close to each other. Similarly, the ratio of X/Z_Horizontal was very close to that of X/Z_Vertical. In addition, it can be observed, that the magnitudes of the ratios of X/Z_Horizontal and X/Z_Vertical were larger than those of V/H_Form Z and V/H_Form X. This suggests that multiple linking introduced more error in the process of maintaining vertical scales across the three forms under Condition 4 and larger examinee group differences between Forms Z and X tended to have similar impact on the performance of the two scale maintenance approaches.

In comparisons of the results under Conditions 1 and 2, the ratios under Conditions 3and 4 tended to be less fluctuating, especially under Condition 3. This indicates that the two scale maintenance approaches showed more consistent performance under Conditions 3 and 4 with a decreasing within-grade variability pattern across grades. Even though the discrepancies between X/Z_Horizontal and X/Z_Vertical under Condition 4 were more noticeable than those of V/H_Form Z and V/H_Form Z, they were still less significant than under Condition 2.

Conditions 5 and 6

Under Conditions 5 and 6, the true means increased from one grade to the next in the same manner as Conditions 1 to 4, but the within-grade variability increased from an SD of 1 for Grade 10 to an SD of 1.4 for Grade 12, a 40% increase across the three grades. The same baseline vertical scale on Form Y was developed under the two

conditions. As with the other two pairs of conditions, the difference between Conditions 5 and 6 existed in examinee sample characteristics across the three forms. Condition 5 represented small group differences across Forms Y, Z, and X; Condition 6 represented large group differences across the three forms.

Adequacy in Recovering the True Proficiency

Correlations

Table 4.16 lists the means and standard deviations of correlations under Conditions 5 and 6 over 100 replications. Average means and standard deviations of correlations across Grades 10-12 are reported in Table 4.16. Similar to what was observed under the other two pairs of conditions, the two scale maintenance approaches yielded very similar means and standard deviations of correlations at each level test on each form under both conditions. The magnitude of the mean correlations under Conditions 5 and 6 was the highest among the three pairs of conditions. In general the mean correlations ranged from 0.910 to 0.946, indicating that the proficiency scales developed by the horizontal approach and by the vertical approach match the true proficiency scales pretty well. However, when single linking (on Form Z) or multiple linking (on Form X) across forms on vertical scales was conducted, some slight differences can be observed between the horizontal approach and the vertical approach.

Figure 4.17 plots the average correlations across grades under the two conditions. By comparing the results between the two scale maintenance approaches on each form under the two conditions with increasing within-grade variability pattern, the vertical

approach again yielded slightly higher average mean correlations across grades than the horizontal approach. The magnitudes of the means and standard deviations of correlations were larger on the interim Form Z when linking the interim Form Z to the base Form Y. When multiple linkings were conducted on Form X, the magnitudes of average correlations were slightly smaller under both conditions, indicating that more drift occurred with multiple linkings across forms. Furthermore, the extent of decrease along with the impact of multiple linking was more evident under Condition 6, indicating that larger examinee group differences across forms introduced more error. For instance, under Condition 5, the average correlations via the horizontal approach decreased from 0.936 on Form Z to 0.930 on Form X; under Condition 6, the average correlations via the horizontal approach decreased from 0.929 on Form Z to 0.916 on Form X.

RMSE and MAD

The means and standard deviations of RMSEs and MADs via the two approaches under Conditions 5 and 6 over 100 replications are reported in Tables 4.17 and 4.18.The means and standard deviations of RMSEs across Grades 10-12 were averaged and labeled as “Average” in the tables.

Figure 4.18 visually demonstrates RMSE trends between the horizontal approach and the vertical approach as grade increases under Conditions 5 and 6. Regardless of approaches and forms, increasing patterns were observed as grade increased under both conditions, which is different from the fluctuating pattern under Conditions 1 and 2 and from the decreasing pattern under Conditions 3 and 4 from low to high grades. It can be

seen under Condition 5, as grade increased, the horizontal approach and the vertical approach yielded almost identical increasing patterns of RMSEs on Forms Z and X with slight discrepancies at Grade 11 on the new Form X.

As seen in Figure 4.18, under Condition 6, increasing patterns were observed as well but the discrepancy between the two scale approaches was larger than what was observed under Condition 5. The vertical approach yielded slightly higher RMSEs on the interim Form Z with single linking. On the new Form X with multiple linkings, the distance of the resulting estimates from the two scale maintenance approaches tended to be larger as grade increased than on the interim Form Z. The extent of increase via the vertical approach tended to be more evident on Form X as grade increased. In general, the horizontal approach showed relatively better performance as grade increased, especially on the new Form X.

The means and standard deviations of MADs via the two approaches under Conditions 5 and 6 over 100 replications are reported in Table 4.18 and are plotted in Figure 4.19. In comparisons of the results of RMSEs, as can be seen in Figures 4.18 and 4.19, the overall magnitudes of MADs were smaller than those of RMSEs under both conditions via the two corresponding scale maintenance approaches on each of the two forms. As expected, similar increasing patterns of MADs to those of corresponding RMSEs were observed as grade increased on Form X under each of the two conditions.

Figure 4.13 plots the average of the mean RMSEs and the mean MADs across grades under the two conditions. Similar to the other two pairs of conditions, the

magnitude of MADs was smaller than that of RMSEs, but both indices showed almost parallel results across conditions. It can be seen in Figure 4.20, within each of the two forms under both conditions, that the horizontal approach yielded smaller RMSE and MAD than the vertical approach. When multiple linkings were involved across Forms X and Z and Y, the RMSEs and the MADs of both the horizontal and the vertical approaches increased under the two conditions. The extent of increase on Form X via the vertical approach was more evident.

Growth Statistics

Table 4.19 and Table 4.20 report the grade-to-grade means, standard deviations, mean differences and effect sizes for the proficiency scales via the two approaches averaged under Conditions 5 and 6 over 100 replications. Figure 4.21 through Figure 4.23 are constructed for visual observations of the results of proficiency estimates via the two scale maintenance approaches across forms under Conditions 5 and 6 (mean estimates, standard deviation estimates, and effect size estimates).

Grade-to-grade growth

The pattern of grade-to-grade growth is represented by means and mean differences between adjacent grades within form. Similar to what was observed under the other pairs of conditions, the two scale maintenance approaches under investigation were able to capture the increasing trend in means across grades under Conditions 5 and 6, regardless of single or multiple linkings. This indicates higher performance at higher grades. Yet the results between the two maintenance approaches provided somewhat

different results on Forms Z and X under each of the two conditions. To examine the differences more closely, Figure 4.21 was constructed to illustrate the differences in the mean estimates on Forms Z and X under each of the two conditions, in the same layout as the figures created for the other two pairs of conditions.

In Figure 4.21, under Conditions 3 on the interim Form Z, the horizontal approach and the vertical approach showed mean estimates very close to the true grade means. On the new Form X with multiple linkings, the horizontal approach yielded almost identical mean estimates to the true means at Grades 10 and 12, but a slight underestimate at Grade 11. The vertical approach underestimated the true means across grades and yielded an almost parallel increasing pattern with the true one, but the distance from the true mean at each grade was larger than the horizontal approach estimate. In terms of the two maintenance approaches, both approaches were able to capture the increasing trend of performance, with larger magnitude of true means on Form X as designed. The horizontal approach tended to perform better than the vertical approach in recovering the true means. The horizontal approach appeared to produce very close growth trends on the new Form X in terms of relative positions of each grade.

As Figure 4.21 depicts, the distance between the true means and the mean estimates via the two scale maintenance approaches was more noticeable under Condition 6 than under Condition 5. Similar to what can be observed on Condition 5 in Figure 4.21, the horizontal approach showed better performance on both forms. Both approaches appeared to underestimate the true means on both forms, and the extent of

underestimation tended to be larger via the vertical approach, especially on the new Form X with multiple linkings. The vertical approach showed better performance in terms of relative positions of each grade on Form X in capturing the similar characteristics of the true growth trend on the new Form X. However, the distances between the true growth trend and the estimated growth trend across grades were larger than the horizontal approach on Form X under Condition 6.

Grade-to-grade Variability

The standard deviation estimates via the two approaches on each of the two forms under Conditions 5 and 6 are reported in Tables 4.19 and 4.20. Under both conditions, the within-grade variability of the true vertical scale increased as grade increased on each of the three forms. With respect to the within-grade variability, underestimation trends can be detected via both approaches across forms under the two conditions, which is similar to what was observed under the other two pairs of conditions. Figure 4.22 provides visual observations about the extent of underestimation of the true standard deviations via the two approaches under Conditions 5 and 6. Increasing patterns of the SD estimates via the two approaches can be observed under Conditions 5 and 6, with overall larger extent of underestimation under Condition 6.

Under both conditions, the horizontal approach and the vertical approach showed similar performance at Grades 10 and 11on the interim Form Z. At Grade 12, the SD estimate via the vertical approach tended to deviate more from the true SD. On the new Form X, the two scale maintenance approaches performed similarly across grades, with

slightly larger underestimation via the horizontal approach at Grade 10. Discrepancies between the resulting SD estimates via the two approaches can be observed at Grade 12 on the interim Form Z and at Grade 10 on the new Form X. In general, multiple linking across three forms tended to have less impact on the resulting SD estimates via both approaches under Condition 5 than under the other two pairs of conditions. Furthermore, the two scale maintenance approaches appeared to show better performance in capturing the true increasing pattern of within-grade variability on the new Form X with multiple linking.

Whether one scale maintenance approach produced a within-grade variability pattern more similar to the true decreasing pattern across grades under the two conditions is not clear. As depicted in Figure 4.22, fluctuation can be observed on the resulting SD estimates via the vertical approach at Grade 12 on Form Z under both conditions. Other than that, the horizontal approach and the vertical approach showed parallel increasing trends across grades on the two forms under Conditions 5 and 6.

Separation of Grade Distributions

The separation of grade distributions can be observed through the effect size estimates between adjacent grade levels. The effect size estimates via the two approaches on each of the two forms under Conditions 5 and 6 are reported in the last section of Tables 4.19 and 4.20. Similar results were observed under Conditions 5 and 6 with regard to the grade-to-grade growth and grade-to-grade variability. Figure 4.22 shows that the corresponding decreasing patterns of effect size estimates via the two scale maintenance

approaches were similar under Conditions 5 and 6. Under Conditions 5 and 6, as reported in Tables 4.19 and 4.20, the true effect sizes showed decreasing trend as grade increased, indicating a decelerated growth from lower to higher grades. Both approaches captured the decreasing pattern of effect size on each form.

On the interim Form Z, the horizontal approach yielded almost identical estimates with the true effect sizes; the vertical approach slightly overestimated the true effect sizes. On the new Form X when multiple linkings were involved, more fluctuation was detected (Figure 4.15). The two approaches tended to underestimate the true effect size at Grade 10/11, but tended to overestimate the true effect size at Grade 11/12. The extent of overestimation via the horizontal approach was more evident than the vertical approach on Form X under Condition 6. In general, under both conditions, the horizontal approach performed better than the vertical approach on the interim Form Z but fluctuations existed at Grade 11/12 on Form X. Thus the vertical approach performed slightly better on Form X in accurately estimating the true effect sizes, but neither of the two approaches resulted in a decreasing pattern similar to the true effect sizes.

Ratios of RMSE and MAD

The results via the two scale maintenance approach in the above two sections are mixed under Conditions 5 and 6. Discrepancies were detected via each approach across forms under each condition, indicating that multiple linkings across three forms had an impact on the resulting indices: correlation, RMSE, MAD, mean, standard deviation, and effect size. It is hard to know whether the discrepancies of the resulting indices between

the horizontal approach and the vertical approach were due to different approaches only, to larger examinee group differences between forms only, or to both approach differences and group differences. In this section, to further investigate under which condition one approach is preferable to the other, Figure 4.24 provides visual observations of the ratios of the average RMSEs and the ratios of the average MADs, in the same layout as Figures 4.8 and 4.16 under the other two pairs of conditions. The average RMSEs and MADs are reported in Tables 4.17 and 4.18. As can be observed in Figure 4.24, under both conditions, the values of the RMSE ratios and the MAD ratios were very consistent for each of the four sets.

Under Condition 5, with comparisons of the Vertical/Horizontal ratios between the two forms, the ratios of V/H_Form X were consistent with that of V/H_Form Z and the values were close to 1, indicating that both approaches showed similar performances with multiple linkings on the new Forms X as well as with single linking on the interim Form Z. It also suggests that multiple linkings across three forms under Condition 5 had almost no impact on the performance of the two scale maintenance approaches. By comparing the X/Z ratios between the two scale maintenance approaches, the ratios of X/Z_Horizontal and X/Z_Vertical were close to each other, which confirmed that the two scale maintenance approaches performed similarly with either single linking or multiple linking. In addition, the magnitudes of the ratios of X/Z_Horizontal and X/Z_Vertical were slightly larger than those of V/H_Form Z and V/H_Form X. However, these

differences were so small that it is hard to tell which scale maintenance approach is better under Condition 5.

Under Condition 6, more discrepancies can be observed between the two approaches. By comparing the Vertical/Horizontal ratios between the two forms, the ratios of V/H_Form Z were higher than the ratios of V/H_Form X, increasing by 0.1. Similarly, the ratios of X/Z_Vertical were higher than those of X/Z_Horizontal, with an increase of 0.1. In addition, the magnitudes of the ratios of X/Z_Horizontal and X/Z_Vertical were larger than those of V/H_Form Z and V/H_Form X. This suggests that multiple linkings introduced more error in the process of maintaining vertical scales across three forms under Condition 6, and larger examinee group differences between Forms Z and X tended to have more impact on the performance of the vertical approach.

Comparisons of the results under Conditions 2 and 4 with large group differences across forms, suggest that multiple linkings tended to have the most impact on the vertical approach under Condition 6, as can be observed from the difference between the ratios of V/H_Form Z and the ratios of V/H_Form X. Also, the discrepancies between the ratios of X/Z_Horizontal and the ratios X/Z_Vertical under Condition 6were more evident than under any of the other two conditions – Conditions 2 and 4.

Table 4.1 Actual Means and SDs for the Simulated Proficiency Distributions under Six Conditions

	Form Y		Form Z		Form X	
	Mean	SD	Mean	SD	Mean	SD
Condition1						
G10	0.001	1.001	0.249	1.001	0.497	0.998
G11	0.665	1.000	0.919	1.000	1.167	1.000
G12	1.027	0.997	1.272	1.000	1.523	1.000
Condition 2						
G10	0.001	1.001	0.499	1.001	0.997	0.998
G11	0.665	1.000	1.169	1.000	1.667	1.000
G12	1.027	0.997	1.522	1.000	2.023	1.000
Condition 3						
G10	0.001	1.001	0.249	1.001	0.497	0.998
G11	0.665	0.921	0.918	0.921	1.167	0.921
G12	1.025	0.629	1.272	0.631	1.523	0.631
Condition 4						
G10	0.001	1.001	0.499	1.001	0.997	0.998
G11	0.665	0.921	1.168	0.921	1.667	0.921
G12	1.025	0.629	1.522	0.631	2.023	0.631
Condition 5						
G10	0.001	1.001	0.249	1.001	0.497	0.998
G11	0.665	1.201	0.919	1.200	1.167	1.200
G12	1.028	1.396	1.272	1.401	1.524	1.400
Condition 6						
G10	0.001	1.001	0.499	1.001	0.997	0.998
G11	0.665	1.201	1.169	1.200	1.667	1.200
G12	1.028	1.396	1.522	1.401	2.024	1.400

Note: In this study, three within-grade variability patterns and two examinee group differences were considered in generating data, which resulted in six conditions. Table 3.6 in Chapter III shows the six conditions of simulated proficiency distributions.

Table 4.2 Raw Mean and Standard Deviation Estimates, Prior to Vertical Scaling

	Form Y		Form Z		Form X	
	Mean	SD	Mean	SD	Mean	SD
Condition 1						
G10	0.002	0.953	0.001	0.950	0.001	0.947
G11	0.002	0.982	0.001	0.969	0.001	0.955
G12	0.003	0.992	0.003	0.980	0.002	0.968
Condition 2						
G10	0.002	0.953	0.001	0.947	-0.001	0.934
G11	0.002	0.979	0.000	0.961	-0.002	0.933
G12	0.003	0.992	0.003	0.966	0.001	0.934
Condition 3						
G10	0.002	0.953	0.001	0.950	0.001	0.947
G11	0.002	0.979	0.000	0.958	0.000	0.944
G12	0.003	0.992	-0.003	0.917	-0.005	0.900
Condition 4						
G10	0.002	0.953	0.001	0.947	-0.001	0.934
G11	0.002	0.979	0.000	0.950	-0.002	0.921
G12	0.003	0.992	-0.005	0.895	0.005	0.858
Condition 5						
G10	0.002	0.953	0.001	0.950	0.001	0.947
G11	0.002	0.979	0.003	0.990	0.003	0.977
G12	0.003	0.992	0.010	1.019	0.008	1.007
Condition 6						
G10	0.002	0.953	0.001	0.947	-0.001	0.934
G11	0.002	0.979	0.003	0.984	0.002	0.959
G12	0.003	0.992	0.013	1.009	0.014	0.985

Note: In this study, three within-grade variability patterns and two examinee group differences were considered in generating data, which resulted in six conditions. Table 3.6 in Chapter III shows the six conditions of simulated proficiency distributions.

Table 4.3 Vertical Linking Functions for Forms Y, Z, and X, Prior to Linking

	Form Y		Form Z		Form X	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Condition 1						
G10/G11	1.073	0.743	0.936	0.692	0.987	0.723
G11/G12	0.986	0.379	0.934	0.374	0.964	0.312
Condition 2						
G10/G11	1.073	0.743	0.938	0.683	1.084	0.667
G11/G12	0.986	0.379	0.910	0.365	0.839	0.399
Condition 3						
G10/G11	0.936	0.668	0.870	0.694	0.917	0.741
G11/G12	0.666	0.332	0.624	0.438	0.877	0.385
Condition 4						
G10/G11	0.936	0.668	0.901	0.632	0.960	0.605
G11/G12	0.666	0.332	0.647	0.396	0.845	0.252
Condition 5						
G10/G11	1.209	0.684	1.135	0.700	1.157	0.620
G11/G12	1.076	0.239	1.091	0.308	1.166	0.406
Condition 6						
G10/G11	1.209	0.684	1.098	0.680	1.157	0.589
G11/G12	1.076	0.239	1.085	0.294	1.162	0.364

Note: In this study, three within-grade variability patterns and two examinee group differences were considered in generating data, which resulted in six conditions. Table 3.6 in Chapter III shows the six conditions of simulated proficiency distributions.

Table 4.4 Linking Functions under the Horizontal Approach

	Form Z_Form Y		Form X_Form Z	
	Slope	Intercept	Slope	Intercept
Condition 1				
G10	0.984	0.251	0.902	0.285
G11	0.831	0.203	0.971	0.217
G12	0.938	0.184	0.964	0.312
Condition 2				
G10	0.975	0.497	0.906	0.535
G11	0.900	0.441	0.959	0.347
G12	1.015	0.555	0.940	0.336
Condition 3				
G10	0.984	0.251	0.902	0.285
G11	0.917	0.238	0.982	0.287
G12	1.086	0.486	0.958	0.441
Condition 4				
G10	0.975	0.497	0.906	0.535
G11	0.927	0.553	0.925	0.492
G12	1.104	0.338	0.996	0.370
Condition 5				
G10	0.984	0.251	0.902	0.285
G11	0.936	0.166	0.982	0.165
G12	1.033	0.192	0.960	0.234
Condition 6				
G10	0.975	0.497	0.906	0.535
G11	0.898	0.354	0.939	0.388
G12	1.011	0.375	0.930	0.396

Note: In this study, three within-grade variability patterns and two examinee group differences were considered in generating data, which resulted in six conditions. Table 3.6 in Chapter III shows the six conditions of simulated proficiency distributions.

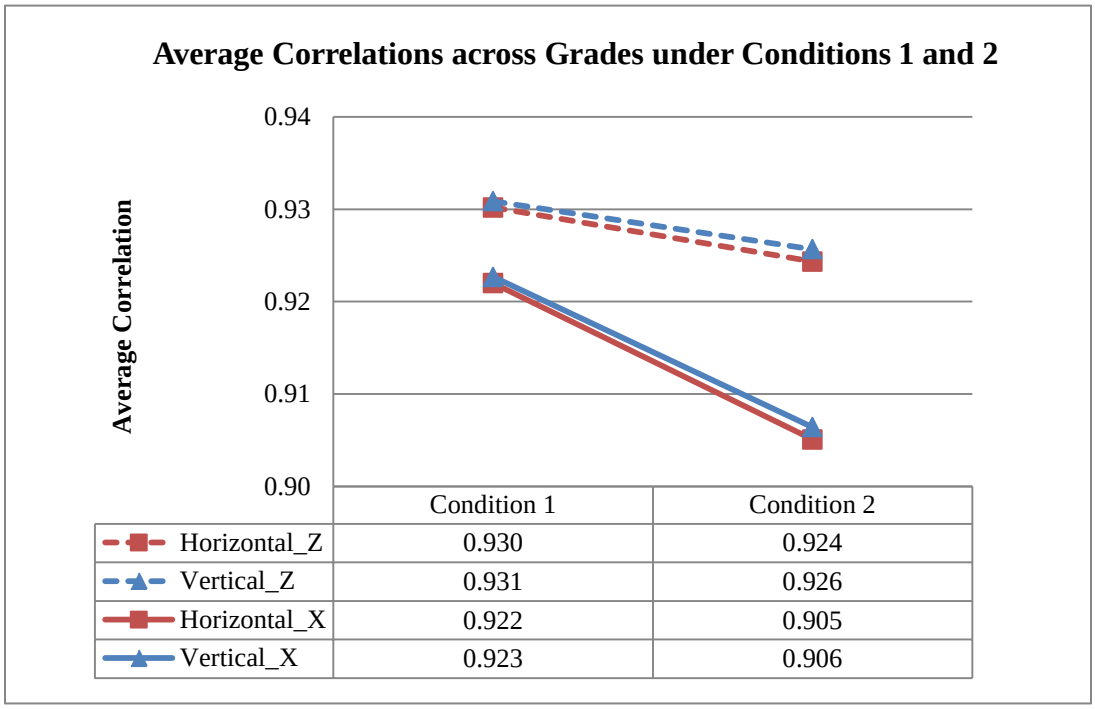
Table 4.5 Linking Functions under the Vertical Approach

	Form Z_ Form Y		Form X_Form Z	
	Slope	Intercept	Slope	Intercept
Condition 1				
All Grades	0.924	0.212	0.943	0.274
Condition 2				
All Grades	0.965	0.499	0.940	0.347
Condition 3				
All Grades	0.985	0.315	0.935	0.330
Condition 4				
All Grades	0.980	0.637	0.924	0.601
Condition 5				
All Grades	0.991	0.197	0.951	0.229
Condition 6				
All Grades	0.972	0.399	0.933	0.437

Note: In this study, three within-grade variability patterns and two examinee group differences were considered in generating data, which resulted in six conditions. Table 3.6 in Chapter III shows the six conditions of simulated proficiency distributions.

Table 4.6 Means and SDs of Correlations between the true proficiencies and the proficiency estimates under Conditions 1 and 2

	Form Z				Form X			
	Horizontal		Vertical		Horizontal		Vertical	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Condition 1								
G10	0.924	0.001	0.924	0.001	0.920	0.004	0.920	0.004
G11	0.934	0.003	0.935	0.004	0.922	0.004	0.923	0.005
G12	0.932	0.002	0.933	0.003	0.924	0.003	0.925	0.004
Average	*0.930*	*0.002*	*0.931*	*0.003*	*0.922*	*0.003*	*0.923*	*0.004*
Condition 2								
G10	0.922	0.002	0.922	0.002	0.910	0.005	0.910	0.005
G11	0.928	0.002	0.930	0.004	0.905	0.005	0.906	0.006
G12	0.924	0.003	0.926	0.003	0.901	0.004	0.904	0.004
Average	*0.925*	*0.002*	*0.926*	*0.003*	*0.905*	*0.004*	*0.906*	*0.005*



	Condition 1	Condition 2
Horizontal_Z	0.930	0.924
Vertical_Z	0.931	0.926
Horizontal_X	0.922	0.905
Vertical_X	0.923	0.906

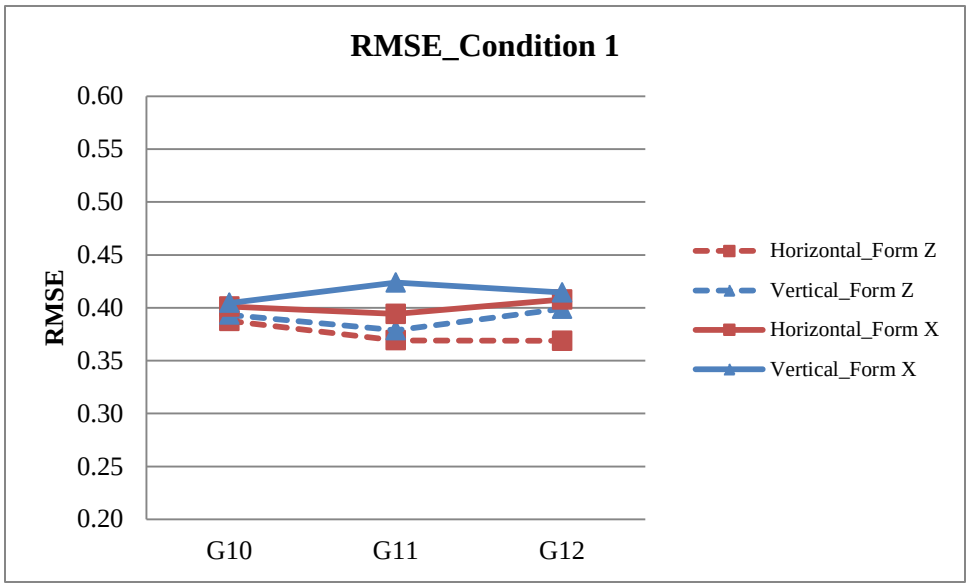
Figure 4.1 Average Correlations between the true proficiencies and the proficiency estimates across Grades under Conditions 1 and 2

Table 4.7 Means and SDs of Root-Mean Square Errors on Conditions 1 and 2

	Form Z				Form X			
	Horizontal		Vertical		Horizontal		Vertical	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Condition 1								
G10	0.388	0.003	0.393	0.007	0.401	0.007	0.404	0.007
G11	0.369	0.005	0.379	0.004	0.394	0.006	0.424	0.018
G12	0.369	0.006	0.399	0.013	0.408	0.011	0.414	0.011
Average	*0.375*	*0.005*	*0.390*	*0.008*	*0.401*	*0.008*	*0.414*	*0.012*
Condition 2								
G10	0.393	0.003	0.393	0.003	0.425	0.005	0.453	0.005
G11	0.385	0.004	0.381	0.004	0.466	0.020	0.480	0.022
G12	0.386	0.010	0.406	0.012	0.477	0.013	0.498	0.015
Average	*0.388*	*0.006*	*0.393*	*0.006*	*0.456*	*0.013*	*0.477*	*0.014*

Table 4.8 Means and SDs of Mean Absolute Differences on Conditions 1 and 2

	Form Z				Form X			
	Horizontal		Vertical		Horizontal		Vertical	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Condition 1								
G10	0.304	0.002	0.308	0.004	0.317	0.008	0.318	0.006
G11	0.288	0.004	0.292	0.004	0.306	0.003	0.327	0.013
G12	0.286	0.004	0.297	0.009	0.320	0.011	0.308	0.008
Average	*0.293*	*0.003*	*0.299*	*0.006*	*0.314*	*0.011*	*0.318*	*0.009*
Condition 2								
G10	0.308	0.002	0.308	0.002	0.334	0.004	0.352	0.003
G11	0.296	0.004	0.294	0.004	0.364	0.017	0.375	0.019
G12	0.295	0.007	0.305	0.008	0.363	0.014	0.389	0.017
Average	*0.299*	*0.004*	*0.302*	*0.004*	*0.354*	*0.012*	*0.372*	*0.013*



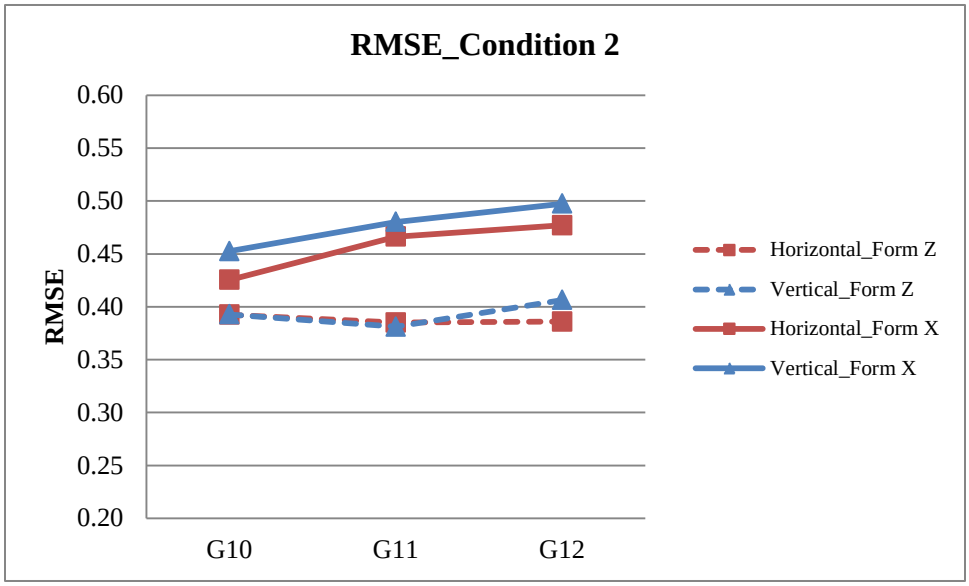
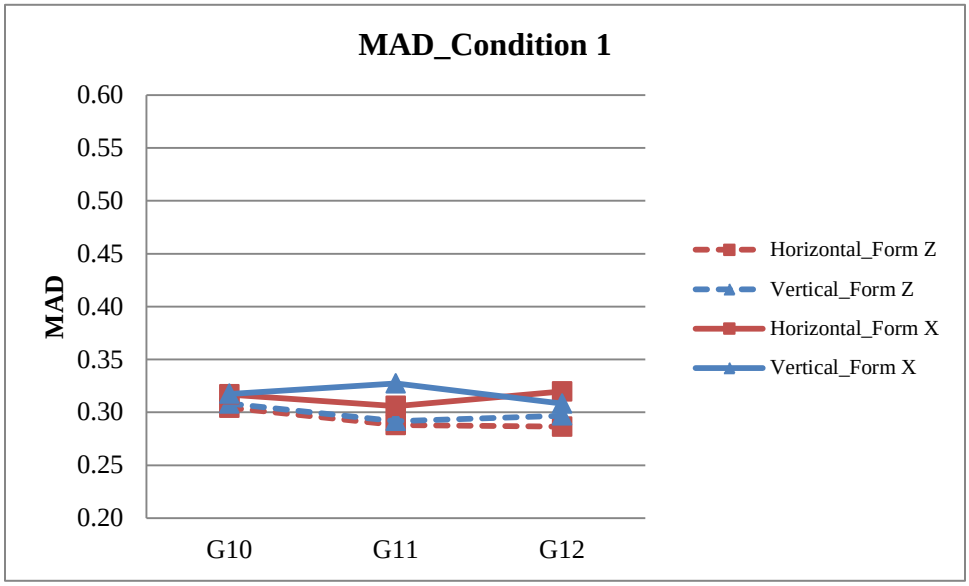


Figure 4.2 RMSEs of the Horizontal Approach and the Vertical Approach on Forms Z and X under Conditions 1 and 2



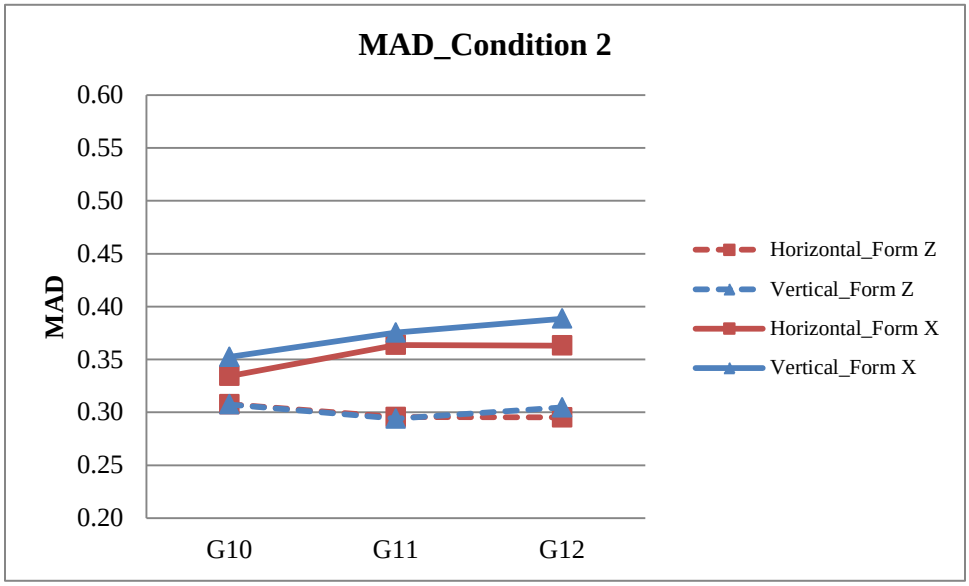


Figure 4.3 MADs of the Horizontal Approach and the Vertical Approach on Forms Z and X under Conditions 1 and 2

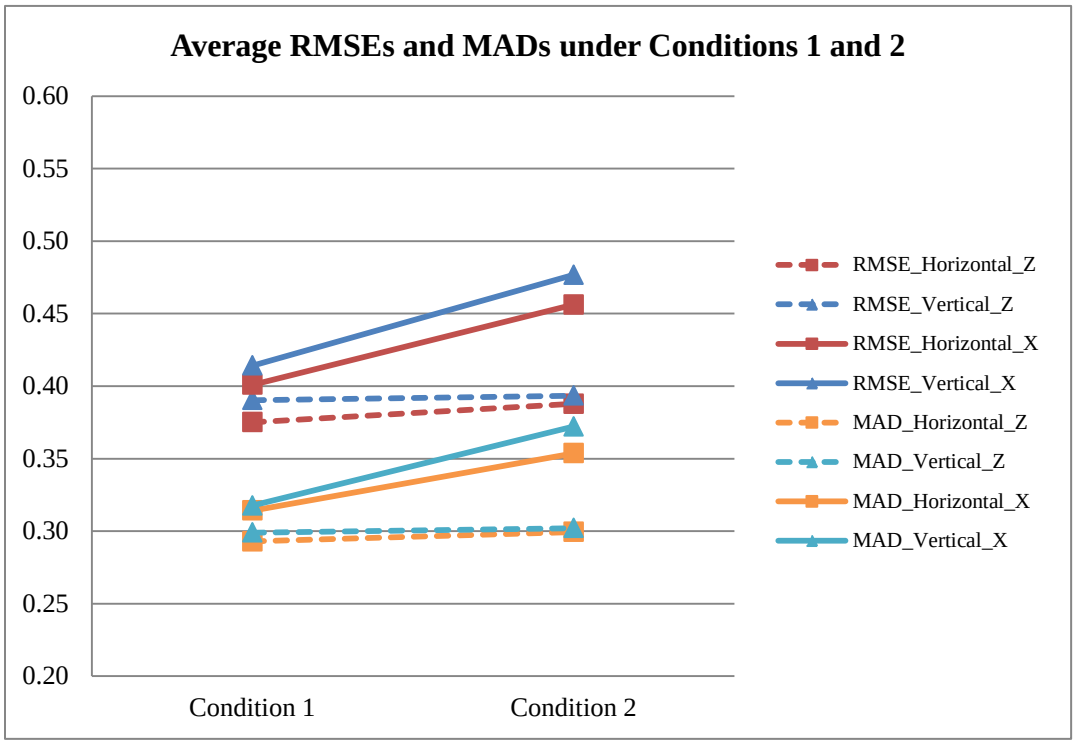


Figure 4.4 Average RMSEs and MADs under Conditions 1 and 2

Table 4.9 Grade-to-grade Means, SDs, Mean Differences, and Effect Sizes under Condition 1

	Form Z			Form X		
Condition 1	TRUE	Horizontal	Vertical	TRUE	Horizontal	Vertical
Grade			Mean			
G10	0.249	0.253	0.214	0.497	0.534	0.468
G11	0.919	0.962	0.852	1.167	1.158	1.098
G12	1.272	1.345	1.175	1.523	1.656	1.366
Grade			SD			
10	1.001	0.966	0.908	0.998	0.884	0.865
11	1.000	0.864	0.838	1.000	0.848	0.843
12	1.000	0.948	0.783	1.000	0.926	0.803
Grade			Mean Difference			
10/11	0.669	0.709	0.639	0.670	0.624	0.630
11/12	0.354	0.383	0.323	0.356	0.498	0.267
Grade			Effect Size			
10/11	0.669	0.774	0.731	0.670	0.720	0.738
11/12	0.354	0.423	0.399	0.356	0.561	0.325

Table 4.10 Grade-to-grade Means, SDs, Mean Differences, and Effect Sizes under Condition 2

	Form Z			Form X		
Condition 2	TRUE	Horizontal	Vertical	TRUE	Horizontal	Vertical
Grade			Mean			
10	0.499	0.499	0.501	0.997	1.021	0.837
11	1.169	1.112	1.158	1.667	1.429	1.441
12	1.522	1.486	1.486	2.023	1.812	1.832
Grade			SD			
10	1.001	0.954	0.944	0.998	0.876	0.899
11	1.000	0.870	0.870	1.000	0.840	0.871
12	1.000	0.899	0.769	1.000	0.845	0.770
Grade			Mean Difference			
10/11	0.669	0.614	0.657	0.670	0.408	0.604
11/12	0.354	0.373	0.328	0.356	0.383	0.390
Grade			Effect Size			
10/11	0.669	0.672	0.724	0.670	0.476	0.683
11/12	0.354	0.422	0.400	0.356	0.454	0.475

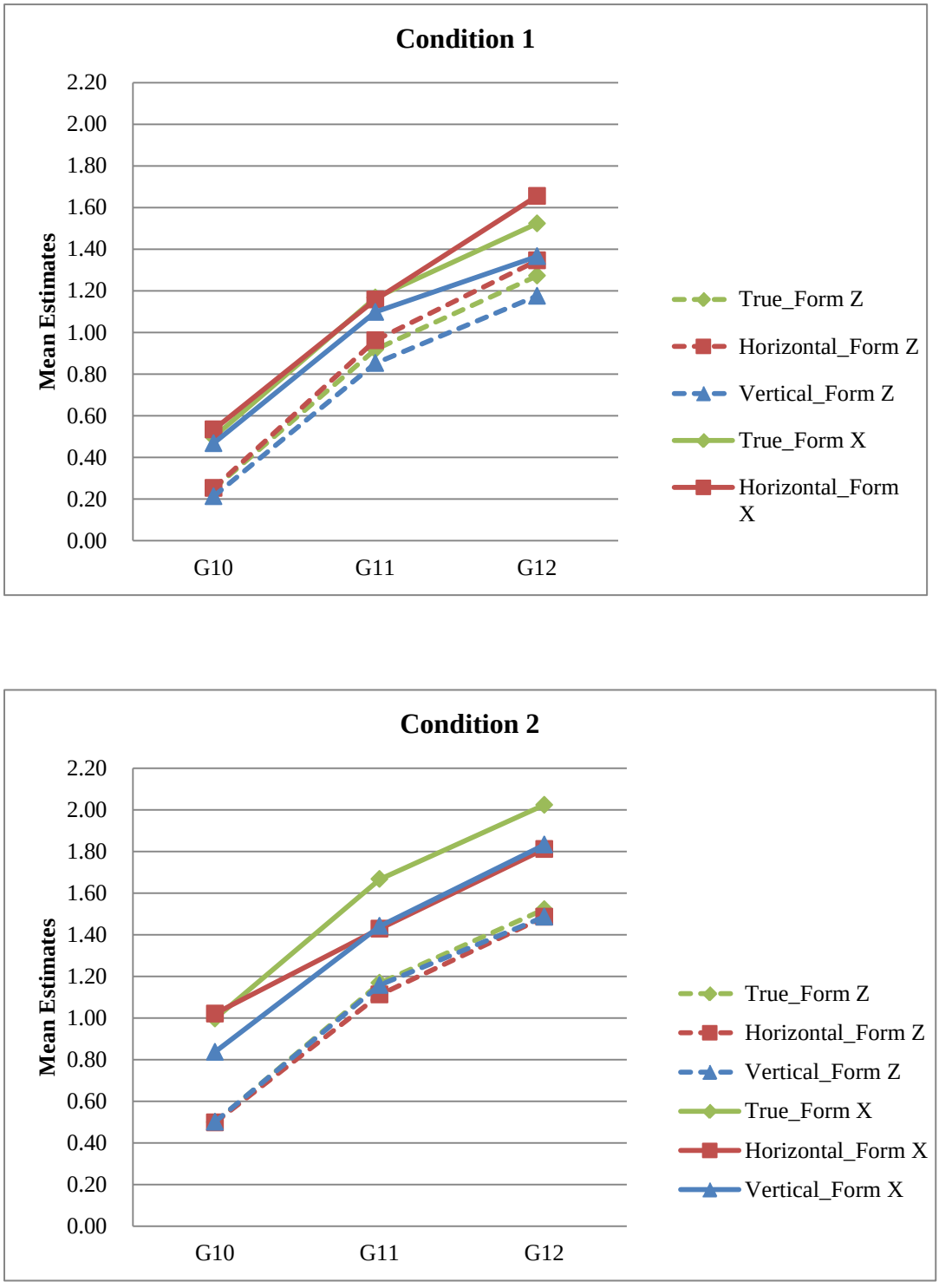
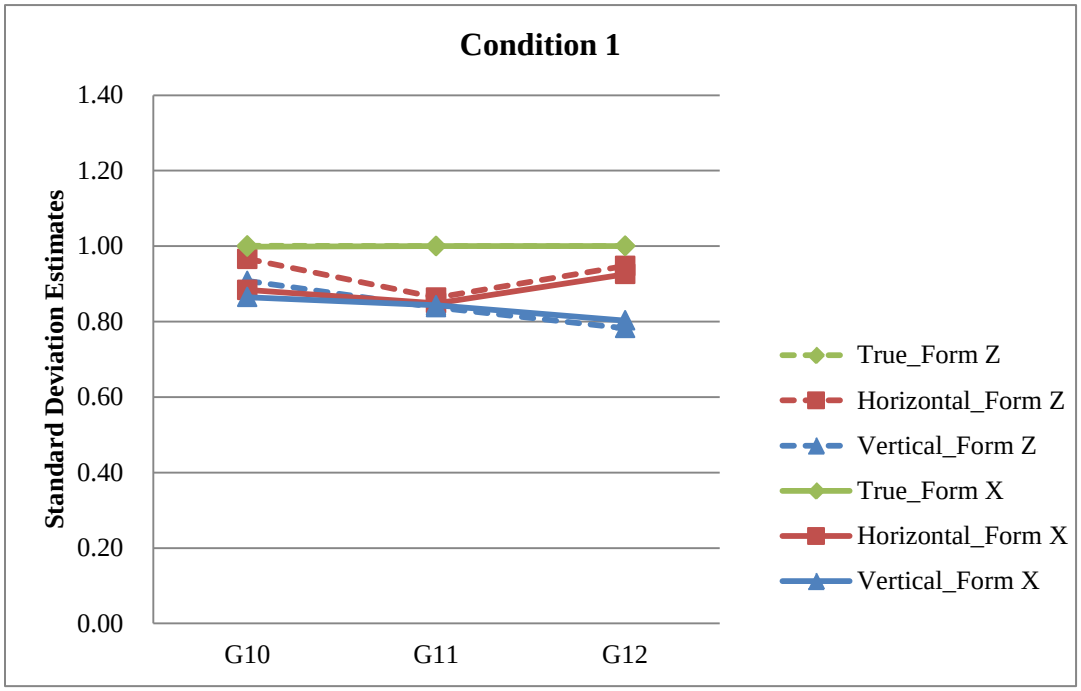


Figure 4.5 Grade-to-grade Mean Estimates under Conditions 1 and 2



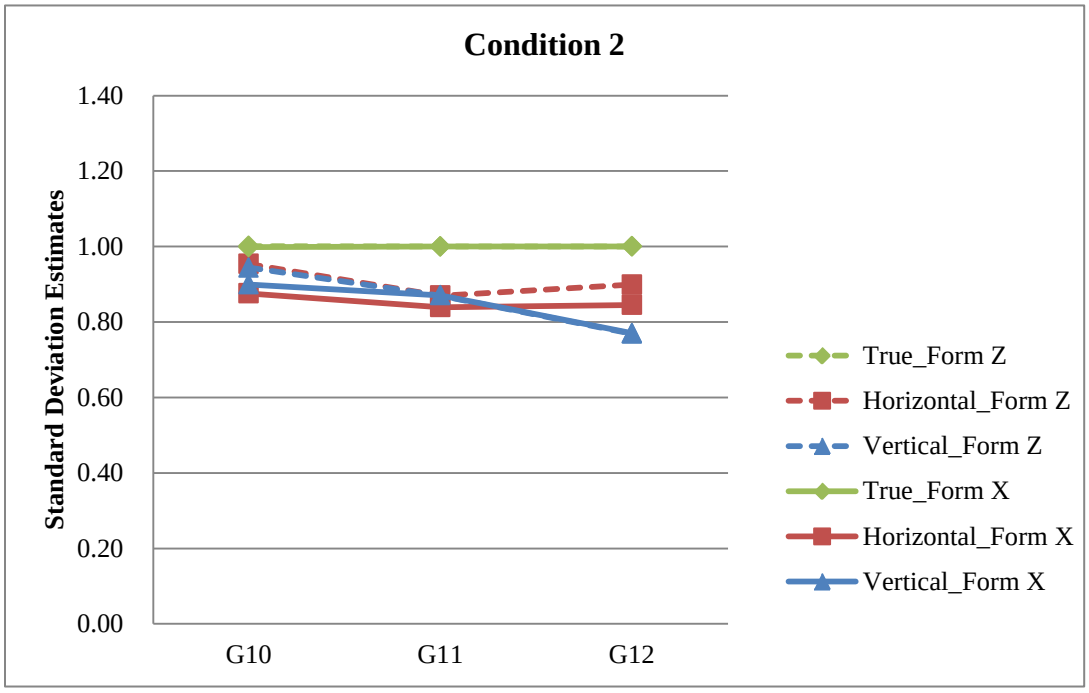
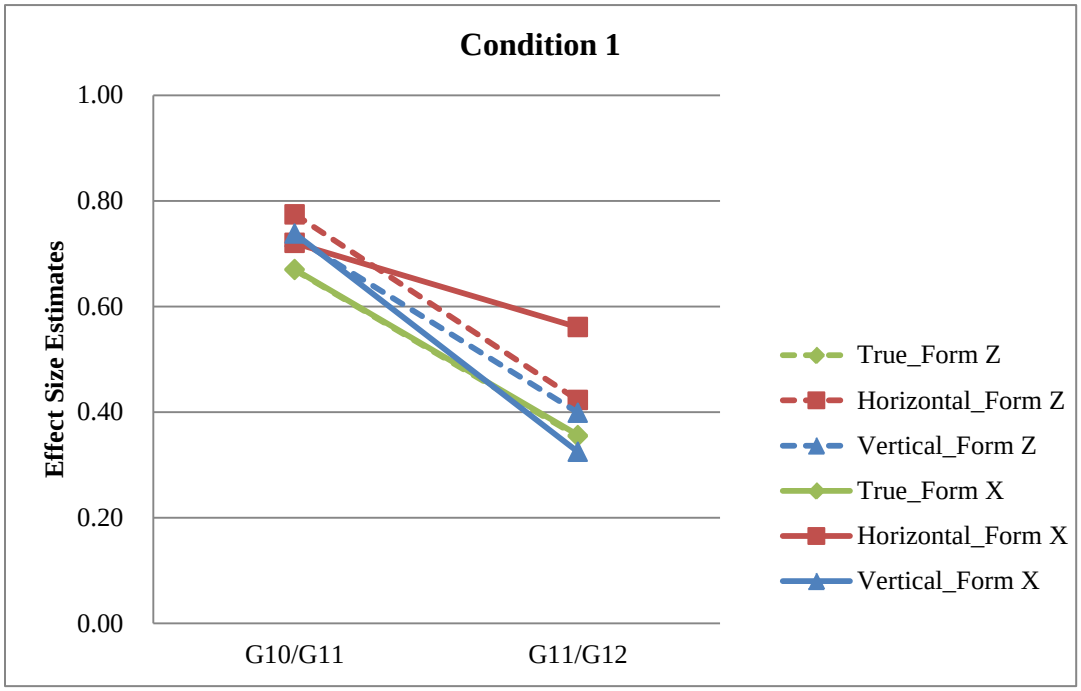


Figure 4.6 Grade-to-grade Standard Deviation Estimates of Ability Estimates under Conditions 1 and 2



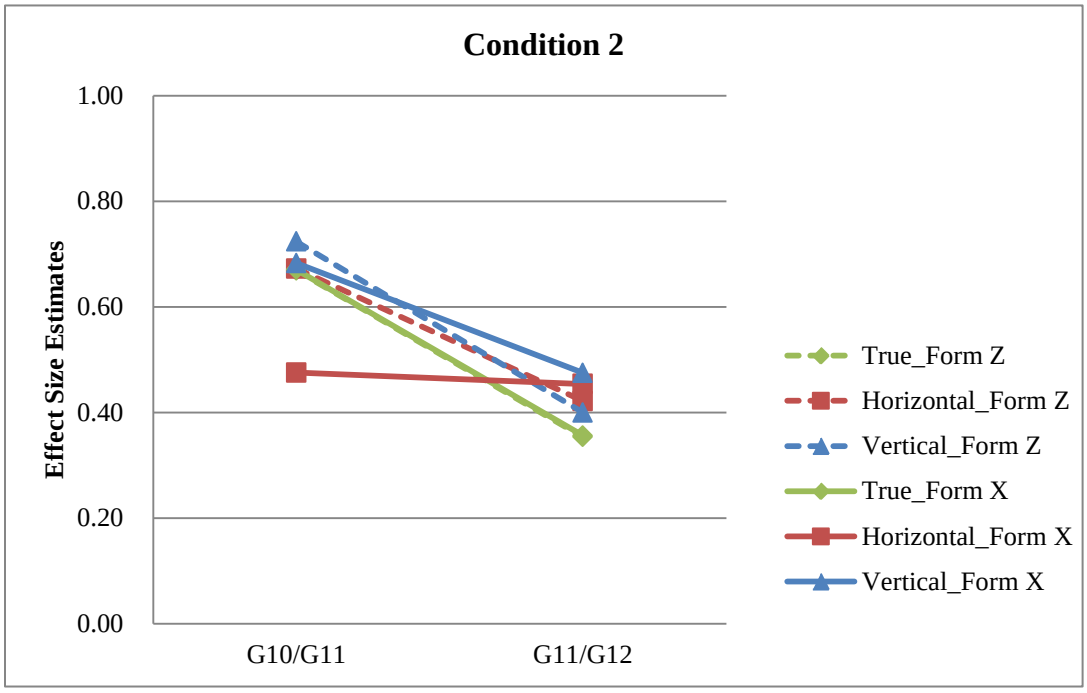
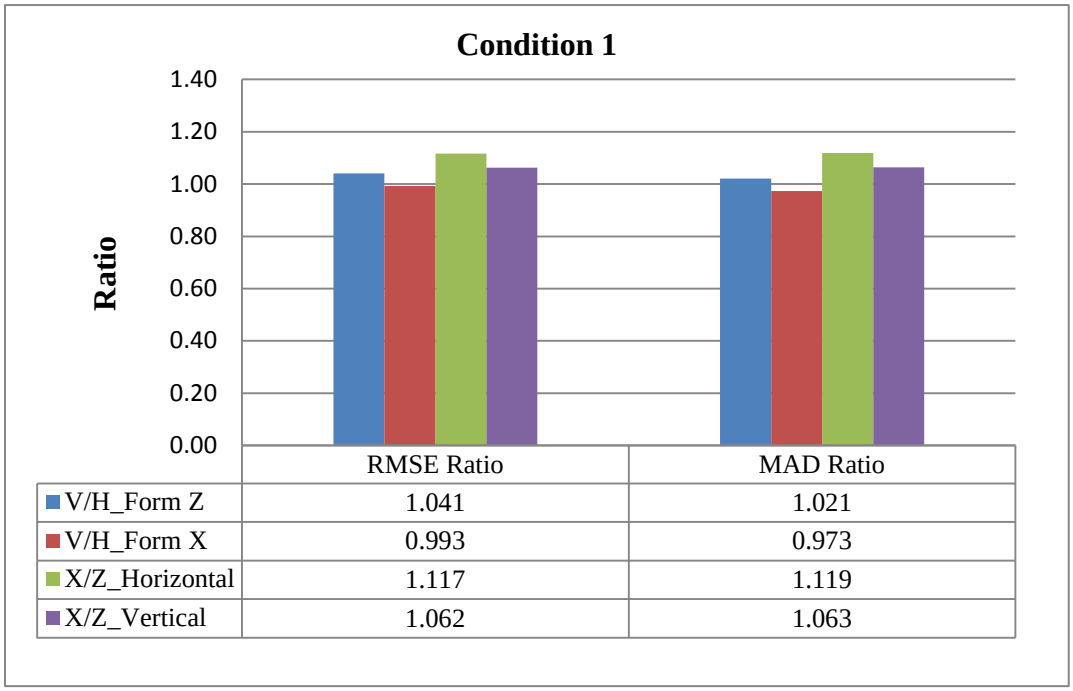
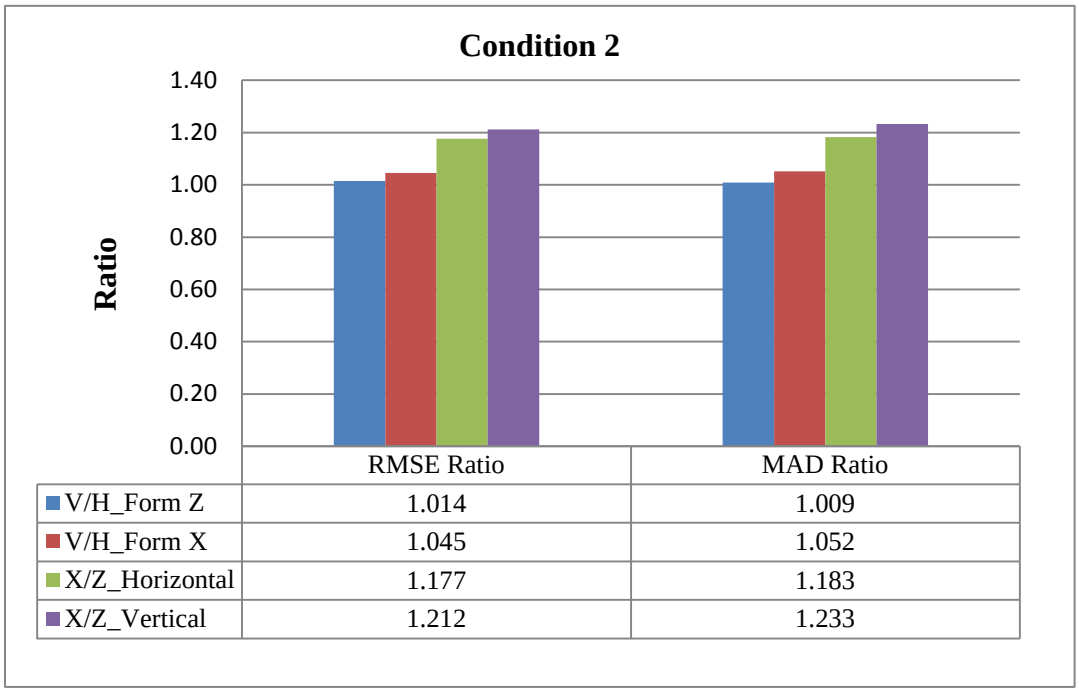


Figure 4.7 Grade-to-grade Effect Size Estimates of Ability Estimates under Conditions 1 and 2



	RMSE Ratio	MAD Ratio
V/H_Form Z	1.041	1.021
V/H_Form X	0.993	0.973
X/Z_Horizontal	1.117	1.119
X/Z_Vertical	1.062	1.063



	RMSE Ratio	MAD Ratio
V/H_Form Z	1.014	1.009
V/H_Form X	1.045	1.052
X/Z_Horizontal	1.177	1.183
X/Z_Vertical	1.212	1.233

Figure 4.8 RMSE Ratios and MAD Ratios between the Scale Maintenance Approaches and between Forms under Conditions 1 and 2

Table 4.11 Means and SDs of Correlations between the true proficiencies and the proficiency estimates under Conditions 3 and 4

	Form Z				Form X			
	Horizontal		Vertical		Horizontal		Vertical	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Condition 3								
G10	0.924	0.001	0.924	0.001	0.920	0.004	0.920	0.004
G11	0.927	0.003	0.929	0.004	0.915	0.003	0.917	0.006
G12	0.913	0.004	0.915	0.003	0.910	0.008	0.915	0.008
Average	*0.921*	*0.003*	*0.923*	*0.003*	*0.915*	*0.005*	*0.917*	*0.006*
Condition 4								
G10	0.923	0.001	0.923	0.001	0.910	0.005	0.910	0.005
G11	0.920	0.003	0.922	0.004	0.907	0.006	0.908	0.006
G12	0.909	0.005	0.910	0.005	0.900	0.005	0.902	0.006
Average	*0.917*	*0.003*	*0.918*	*0.003*	*0.906*	*0.005*	*0.907*	*0.006*

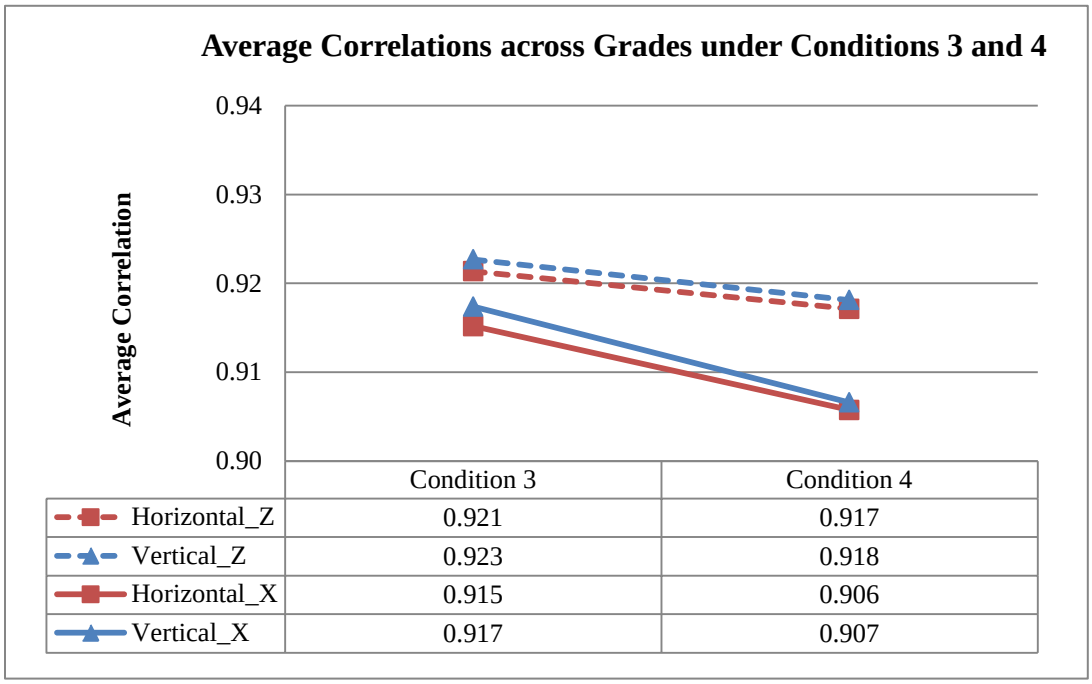


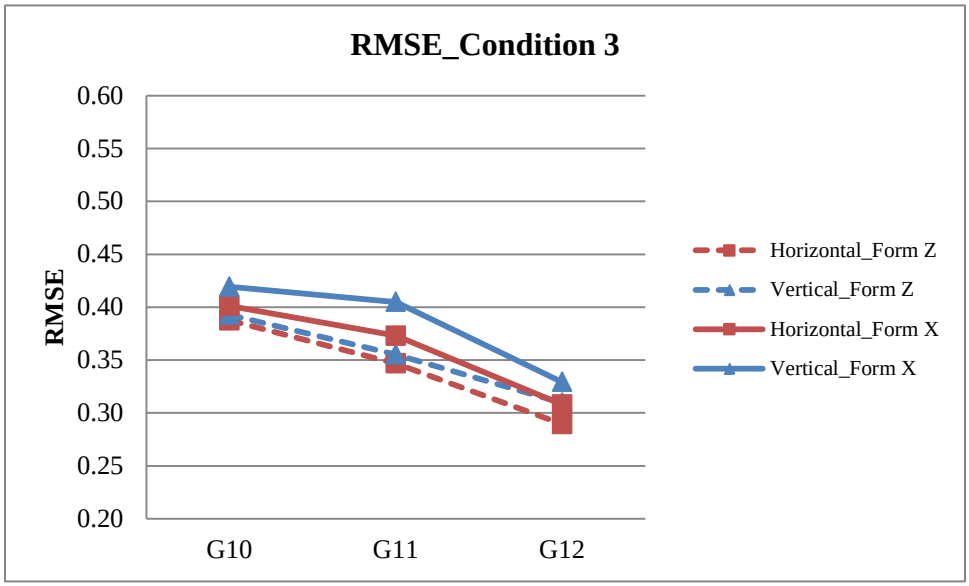
Figure 4.9 Average Correlations between the true proficiencies and the proficiency estimates across Grades under Conditions 3 and 4

Table 4.12 Means and SDs of Root-Mean Square Errors on Conditions 3 and 4

	Form Z				Form X			
	Horizontal		Vertical		Horizontal		Vertical	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Condition 3								
G10	0.388	0.003	0.393	0.003	0.401	0.007	0.419	0.009
G11	0.347	0.004	0.355	0.006	0.373	0.007	0.405	0.011
G12	0.289	0.004	0.310	0.005	0.308	0.006	0.329	0.008
Average	*0.341*	*0.004*	*0.353*	*0.005*	*0.361*	*0.007*	*0.384*	*0.009*
Condition 4								
G10	0.392	0.004	0.413	0.009	0.427	0.005	0.474	0.011
G11	0.363	0.003	0.371	0.005	0.418	0.013	0.422	0.011
G12	0.315	0.004	0.318	0.001	0.390	0.004	0.366	0.007
Average	*0.357*	*0.003*	*0.368*	*0.005*	*0.411*	*0.008*	*0.421*	*0.010*

Table 4.13 Means and SDs of Mean Absolute Differences on Conditions 3 and 4

	Form Z				Form X			
	Horizontal		Vertical		Horizontal		Vertical	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Condition 3								
G10	0.304	0.002	0.308	0.003	0.317	0.007	0.333	0.009
G11	0.271	0.003	0.281	0.004	0.291	0.005	0.321	0.012
G12	0.228	0.003	0.248	0.004	0.242	0.005	0.264	0.008
Average	*0.268*	*0.003*	*0.279*	*0.004*	*0.283*	*0.006*	*0.306*	*0.010*
Condition 4								
G10	0.308	0.002	0.326	0.007	0.336	0.004	0.381	0.008
G11	0.282	0.004	0.291	0.004	0.322	0.011	0.332	0.010
G12	0.247	0.005	0.252	0.002	0.309	0.005	0.290	0.007
Average	*0.279*	*0.003*	*0.290*	*0.004*	*0.322*	*0.006*	*0.334*	*0.009*



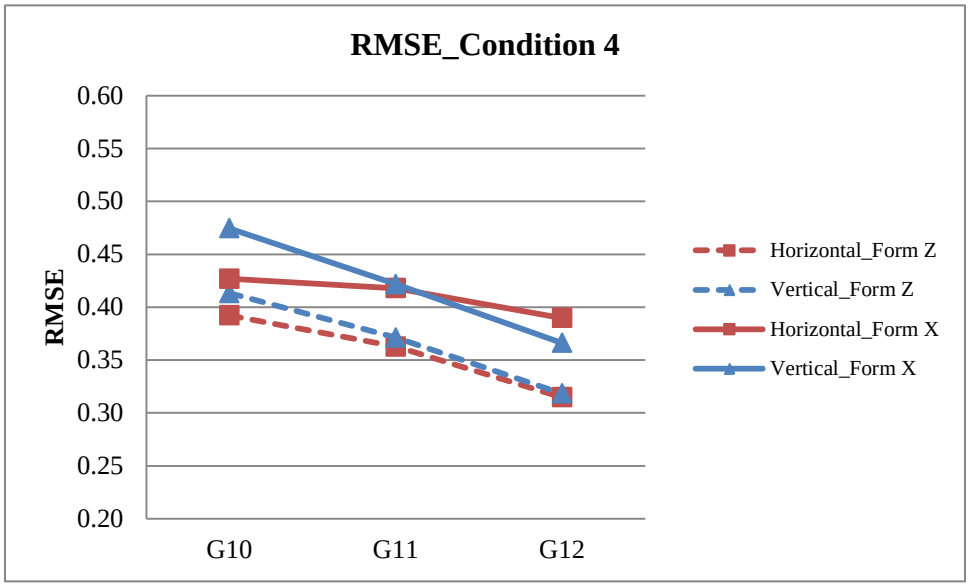
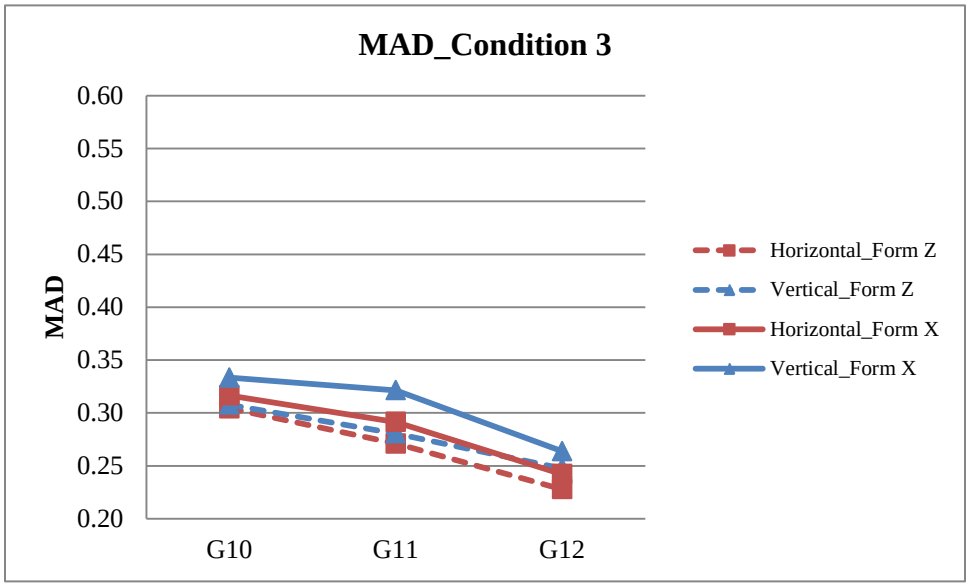


Figure 4.10 RMSEs of the Horizontal Approach and the Vertical Approach on Forms Z and X under Conditions 3 and 4



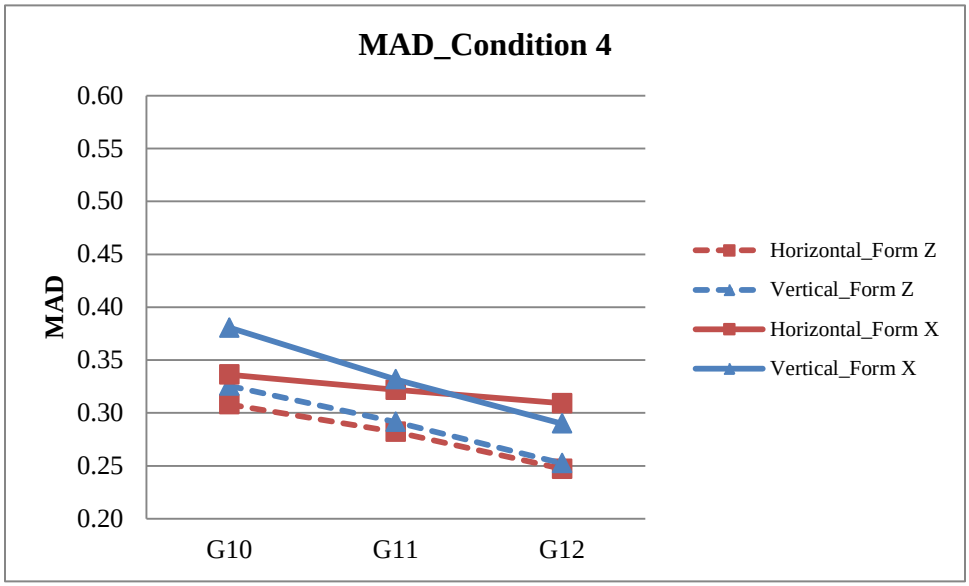


Figure 4.11 MADs of the Horizontal Approach and the Vertical Approach on Forms Z and X under Conditions 3 and 4

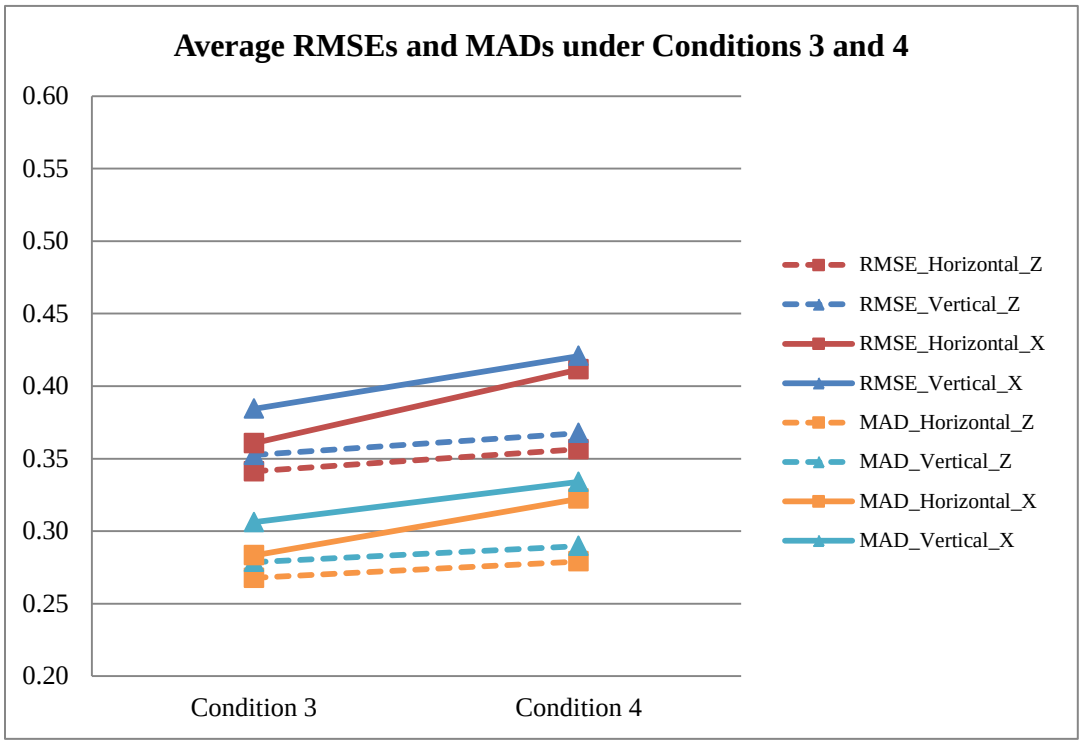


Figure 4.12 Average RMSEs and MADs under Conditions 3 and 4

Table 4.14 Grade-to-grade Means, SDs, Mean Differences, and Effect Sizes under Condition 3

	Form Z			Form X		
Condition 3	TRUE	Horizontal	Vertical	TRUE	Horizontal	Vertical
Grade			Mean			
10	0.249	0.253	0.317	0.497	0.534	0.643
11	0.918	0.892	1.000	1.167	1.135	1.321
12	1.272	1.282	1.376	1.523	1.577	1.645
Grade			SD			
10	1.001	0.963	0.964	0.998	0.881	0.914
11	0.921	0.822	0.821	0.921	0.772	0.775
12	0.631	0.639	0.505	0.631	0.584	0.667
Grade			Mean Difference			
10/11	0.669	0.639	0.683	0.670	0.602	0.678
11/12	0.354	0.391	0.376	0.356	0.442	0.324
Grade			Effect Size			
10/11	0.696	0.714	0.762	0.697	0.727	0.800
11/12	0.448	0.531	0.551	0.451	0.646	0.448

Table 4.15 Grade-to-grade Means, SDs, Mean Differences, and Effect Sizes under Condition 4

	Form Z			Form X		
Condition 4	TRUE	Horizontal	Vertical	TRUE	Horizontal	Vertical
Grade			Mean			
10	0.499	0.499	0.638	0.997	1.021	1.228
11	1.168	1.185	1.256	1.667	1.608	1.770
12	1.522	1.561	1.604	2.023	1.848	2.074
Grade			SD			
10	1.001	0.954	0.959	0.998	0.876	0.899
11	0.921	0.824	0.839	0.921	0.719	0.864
12	0.631	0.633	0.526	0.631	0.588	0.565
Grade			Mean Difference			
10/11	0.669	0.687	0.617	0.670	0.587	0.542
11/12	0.354	0.376	0.349	0.356	0.239	0.304
Grade			Effect Size			
10/11	0.696	0.770	0.685	0.697	0.733	0.615
11/12	0.448	0.511	0.498	0.451	0.365	0.416

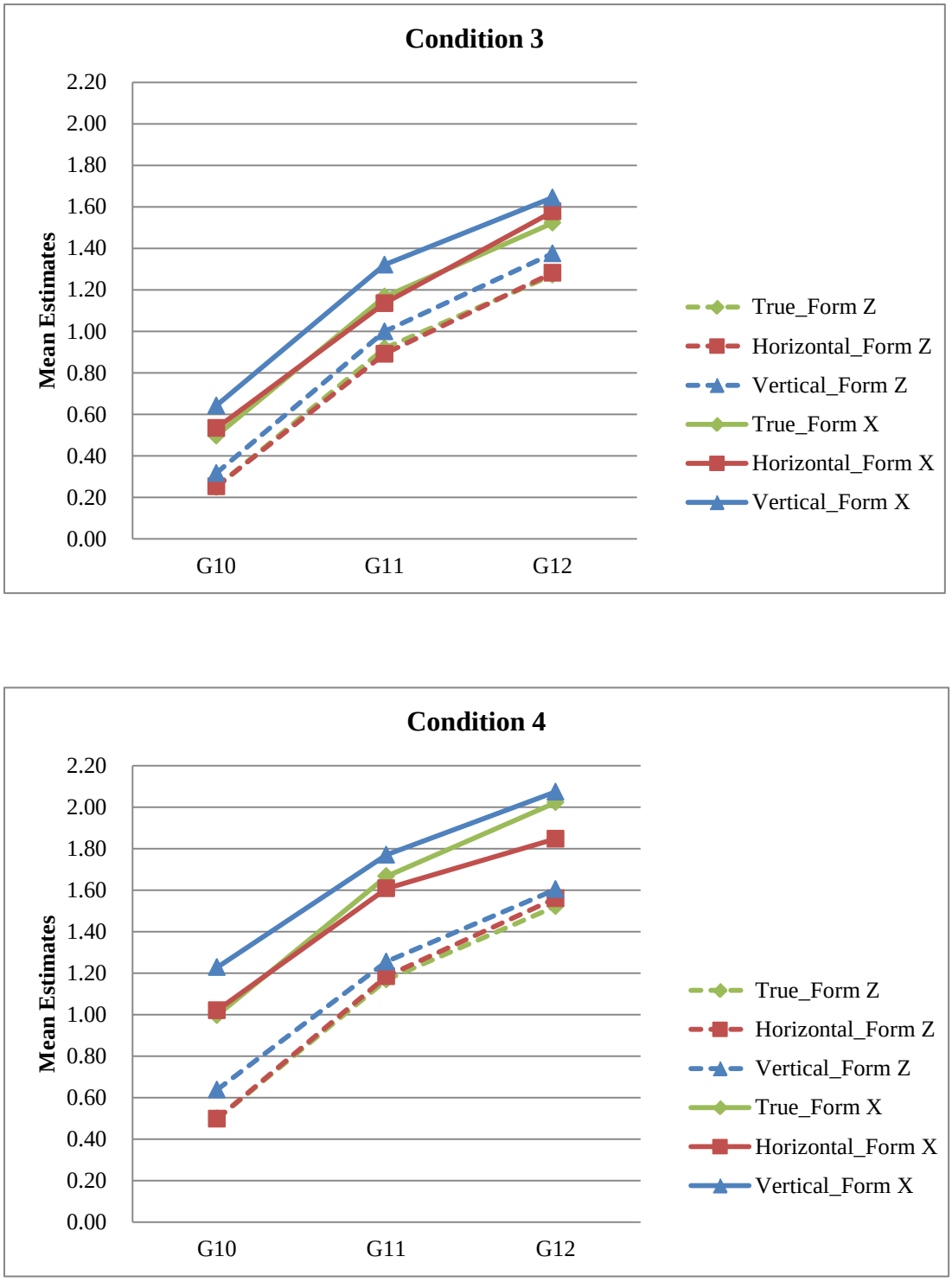


Figure 4.13 Grade-to-grade Mean Estimates under Conditions 3 and 4

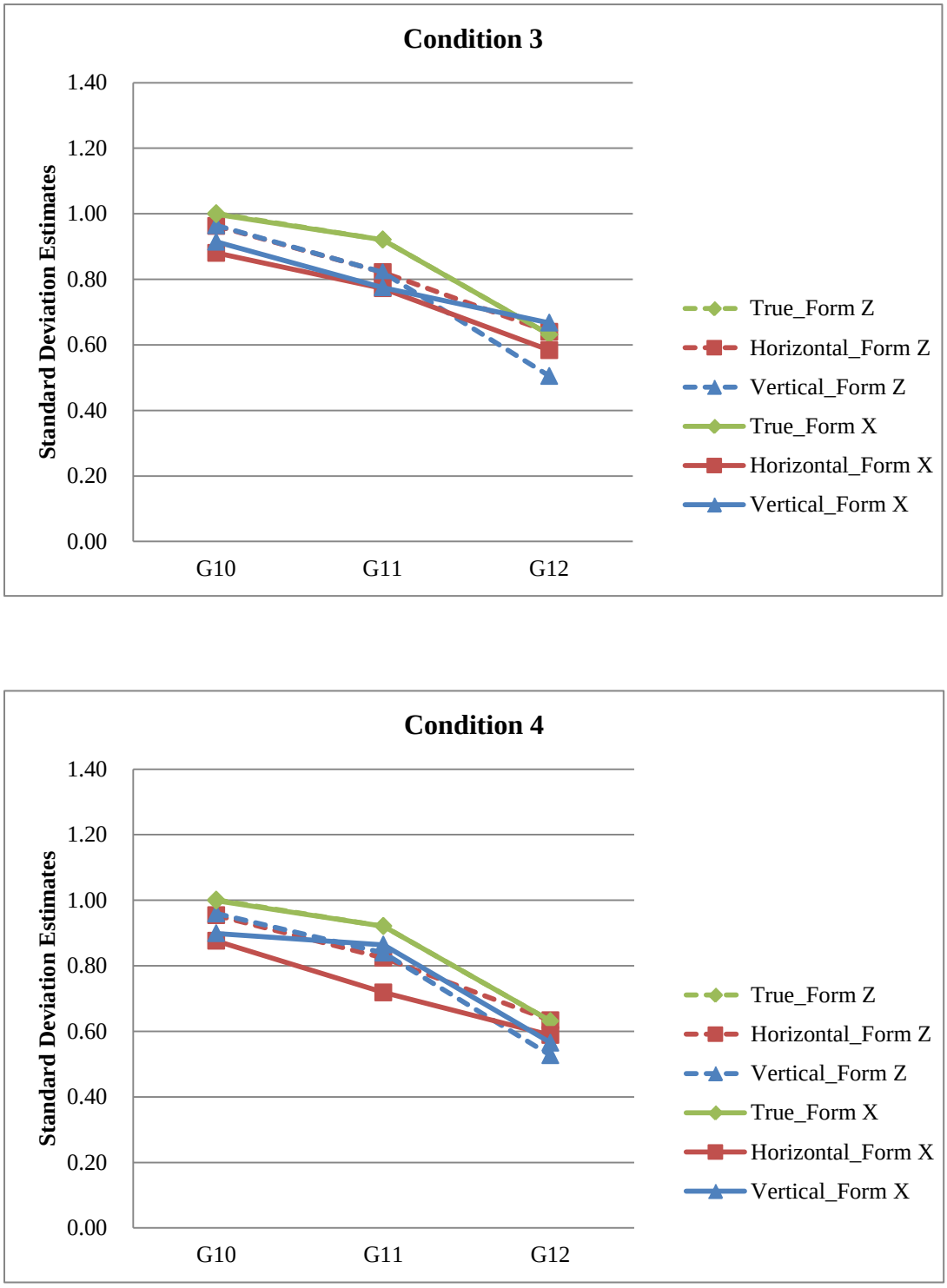
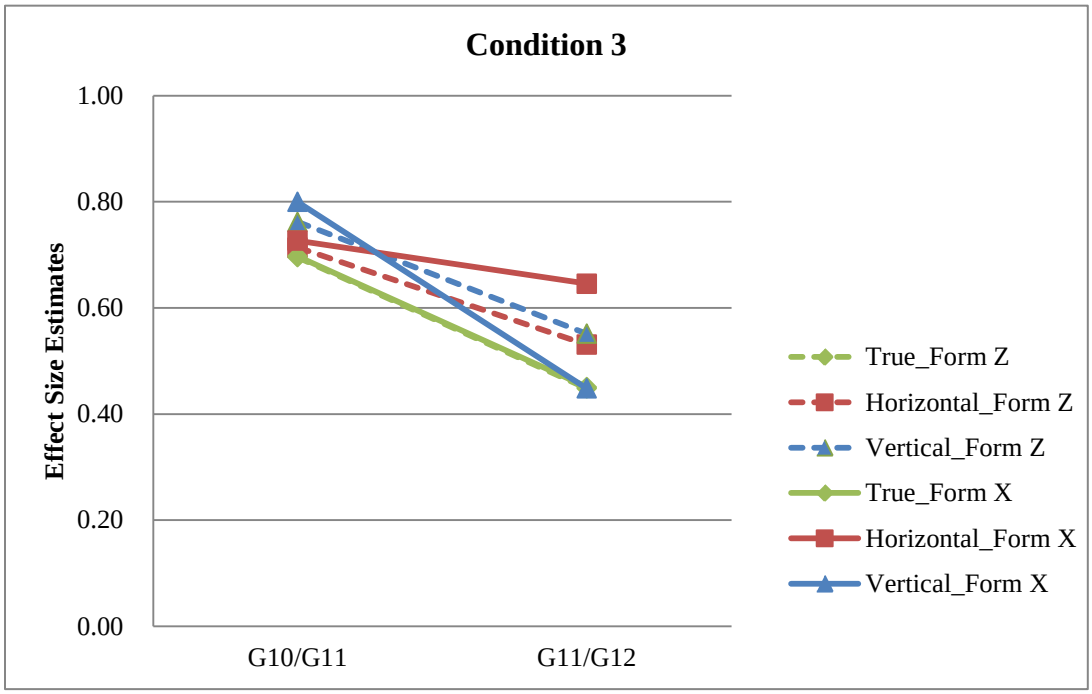


Figure 4.14 Grade-to-grade Standard Deviation Estimates of Ability Estimates under Conditions 3 and 4



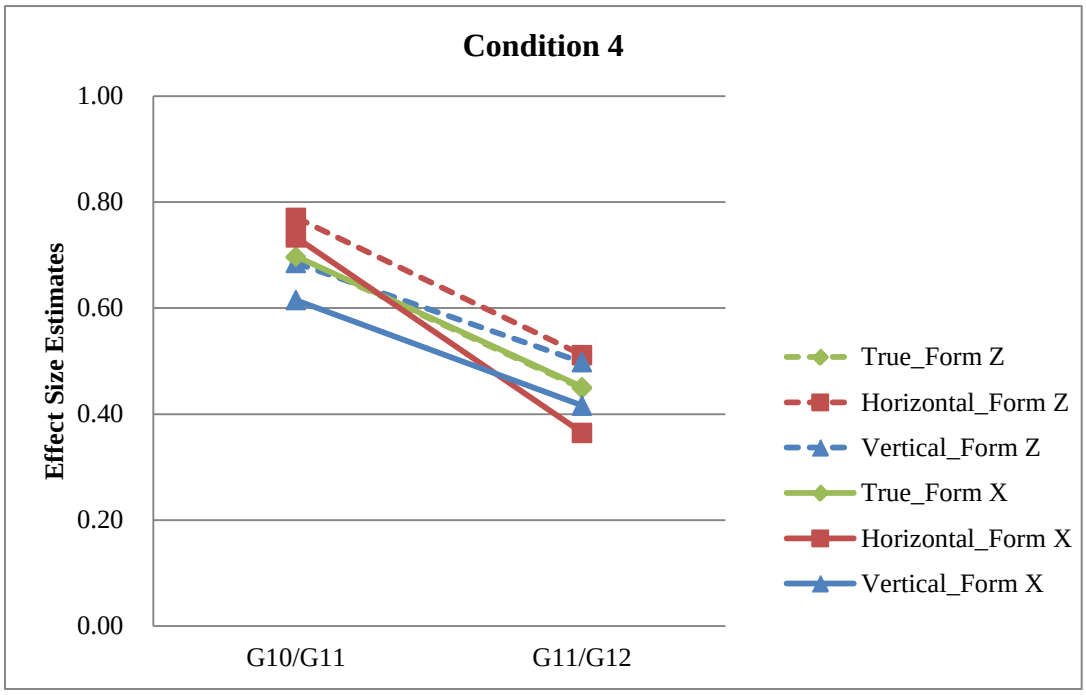
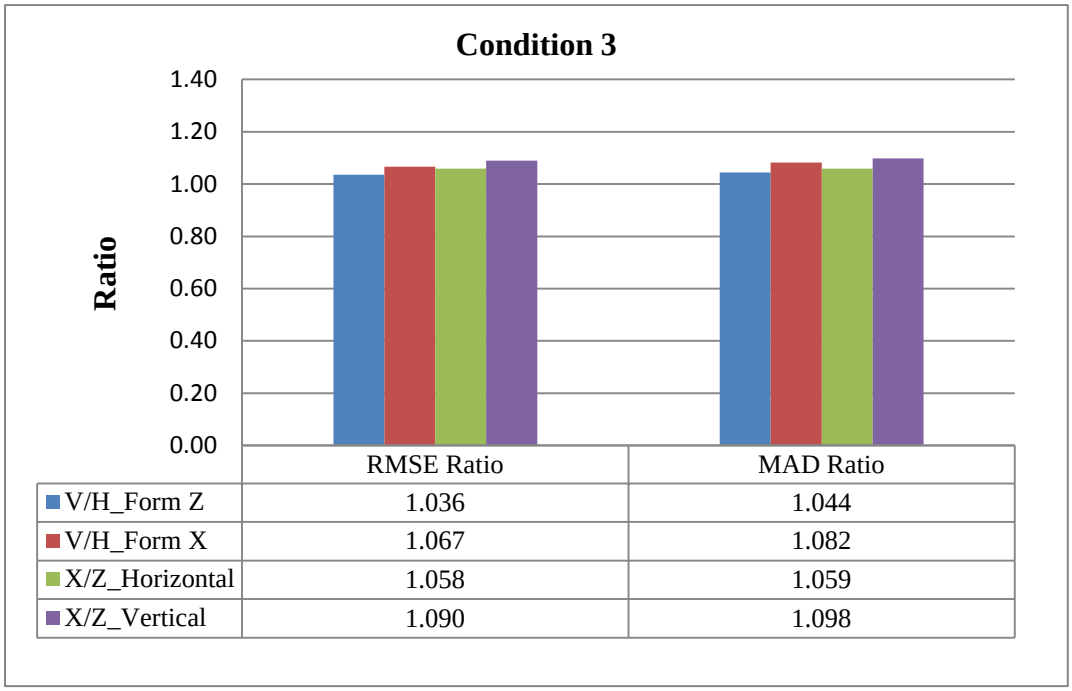
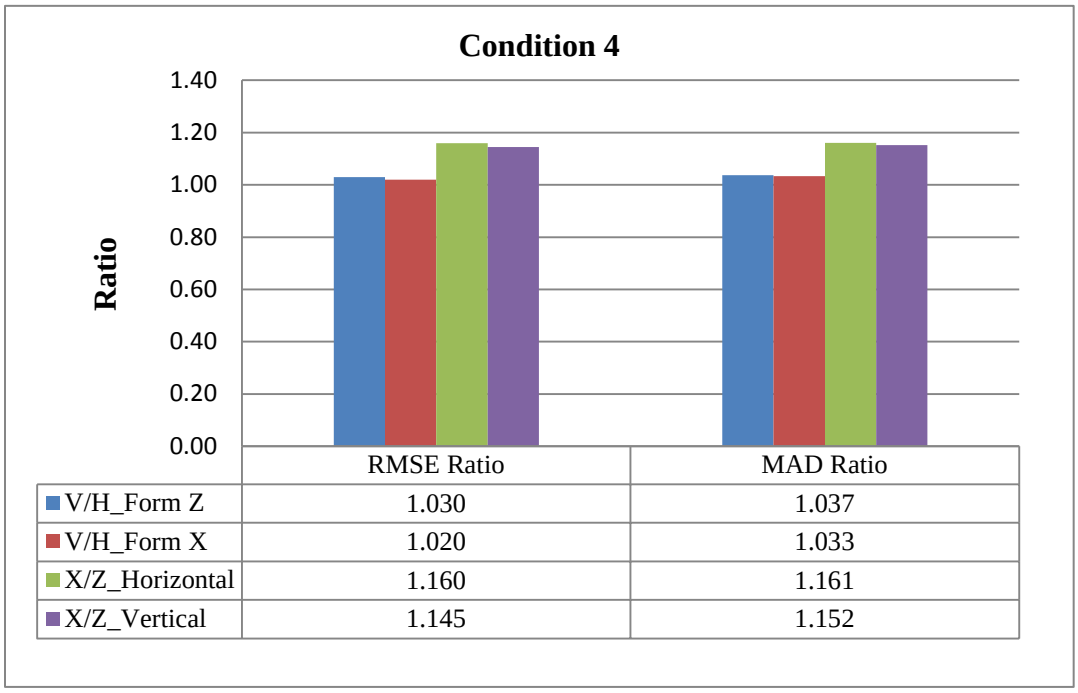


Figure 4.15 Grade-to-grade Effect Size Estimates of Ability Estimates under Conditions 3 and 4



	RMSE Ratio	MAD Ratio
V/H_Form Z	1.036	1.044
V/H_Form X	1.067	1.082
X/Z_Horizontal	1.058	1.059
X/Z_Vertical	1.090	1.098



	RMSE Ratio	MAD Ratio
V/H_Form Z	1.030	1.037
V/H_Form X	1.020	1.033
X/Z_Horizontal	1.160	1.161
X/Z_Vertical	1.145	1.152

Figure 4.16 RMSE Ratios and MAD Ratios between the Scale Maintenance Approaches and between Forms under Conditions 3 and 4

Table 4.16 Means and SDs of Correlations between the true proficiencies and the proficiency estimates under Conditions 5 and 6

	Form Z				Form X			
	Horizontal		Vertical		Horizontal		Vertical	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Condition 5								
G10	0.924	0.001	0.924	0.001	0.920	0.004	0.920	0.004
G11	0.942	0.002	0.946	0.002	0.936	0.003	0.939	0.003
G12	0.941	0.001	0.944	0.001	0.934	0.003	0.936	0.002
Average	*0.936*	*0.001*	*0.938*	*0.001*	*0.930*	*0.003*	*0.932*	*0.003*
Condition 6								
G10	0.922	0.002	0.922	0.002	0.910	0.005	0.910	0.005
G11	0.934	0.002	0.938	0.002	0.921	0.005	0.924	0.005
G12	0.932	0.002	0.935	0.003	0.916	0.003	0.918	0.004
Average	*0.929*	*0.002*	*0.932*	*0.002*	*0.916*	*0.004*	*0.917*	*0.005*

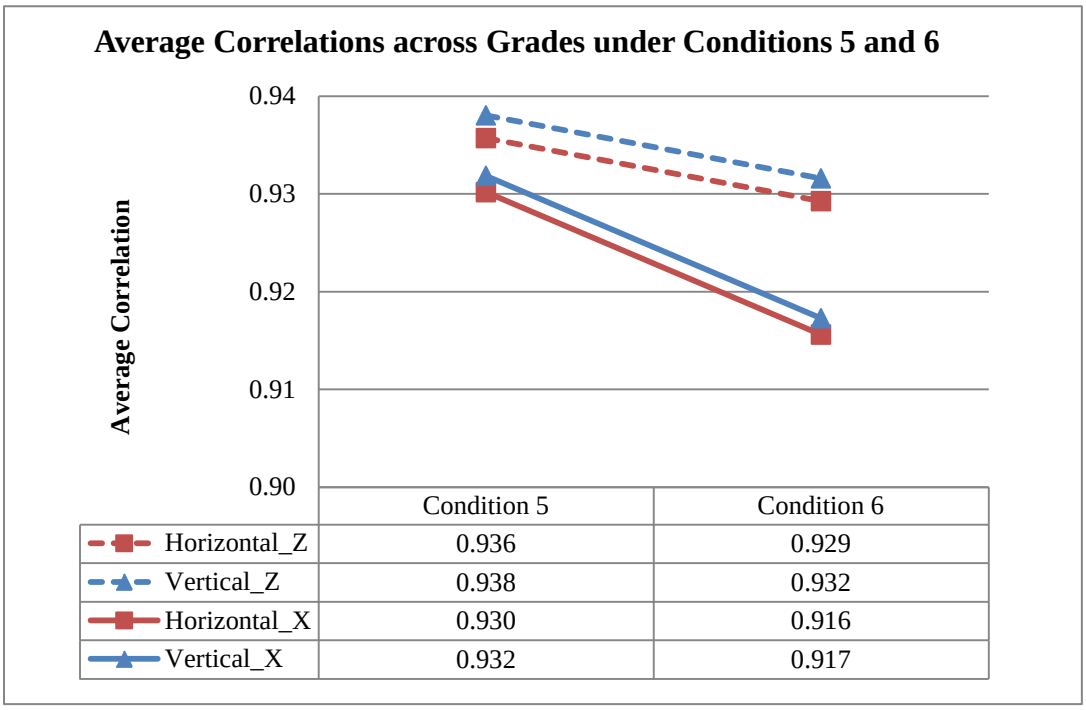


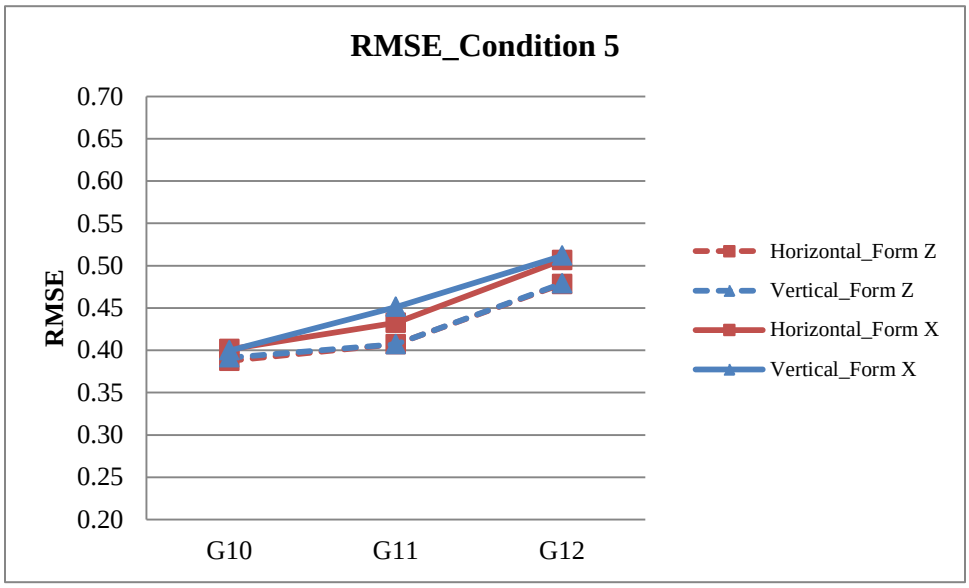
Figure 4.17 Average Correlations between the true proficiencies and the proficiency estimates across Grades under Conditions 5 and 6

Table 4.17 Means and SDs of Root-Mean Square Errors on Conditions 5 and 6

	Form Z				Form X			
	Horizontal		Vertical		Horizontal		Vertical	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Condition 5								
G10	0.388	0.003	0.391	0.006	0.401	0.007	0.400	0.007
G11	0.407	0.002	0.407	0.002	0.432	0.016	0.451	0.021
G12	0.478	0.008	0.479	0.011	0.507	0.013	0.512	0.013
Average	*0.424*	*0.004*	*0.426*	*0.006*	*0.447*	*0.012*	*0.454*	*0.014*
Condition 6								
G10	0.393	0.003	0.405	0.006	0.425	0.005	0.456	0.005
G11	0.428	0.004	0.441	0.005	0.497	0.019	0.562	0.029
G12	0.499	0.010	0.534	0.017	0.577	0.015	0.685	0.025
Average	*0.440*	*0.006*	*0.460*	*0.009*	*0.500*	*0.013*	*0.568*	*0.020*

Table 4.18 Means and SDs of Mean Absolute Differences on Conditions 5 and 6

	Form Z				Form X			
	Horizontal		Vertical		Horizontal		Vertical	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Condition 5								
G10	0.304	0.002	0.308	0.004	0.317	0.007	0.315	0.006
G11	0.317	0.004	0.316	0.004	0.334	0.010	0.348	0.015
G12	0.368	0.005	0.359	0.007	0.382	0.010	0.380	0.009
Average	*0.330*	*0.004*	*0.328*	*0.005*	*0.344*	*0.009*	*0.348*	*0.010*
Condition 6								
G10	0.308	0.002	0.319	0.004	0.334	0.004	0.356	0.003
G11	0.326	0.004	0.334	0.004	0.372	0.013	0.429	0.025
G12	0.377	0.005	0.388	0.011	0.425	0.013	0.507	0.023
Average	*0.337*	*0.004*	*0.347*	*0.006*	*0.377*	*0.010*	*0.431*	*0.017*



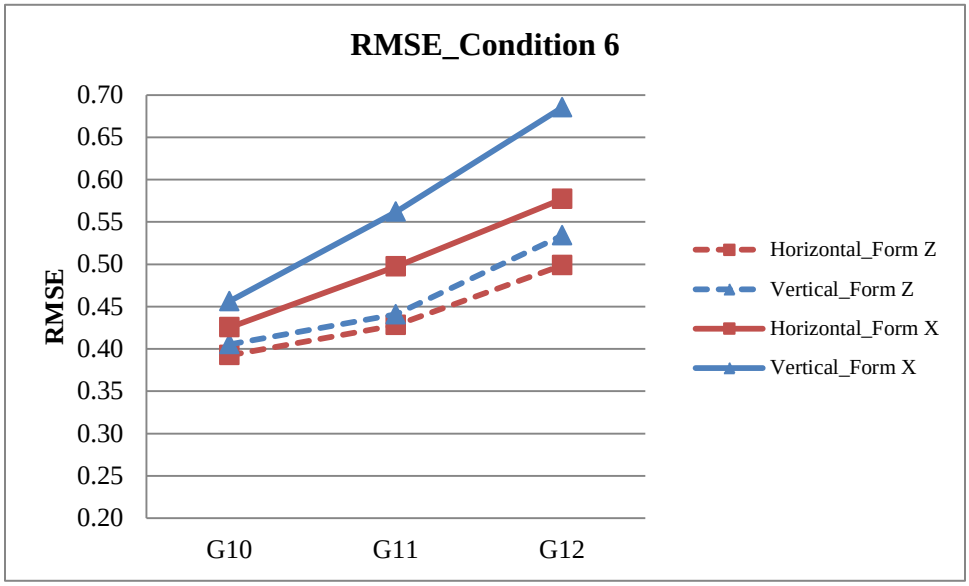
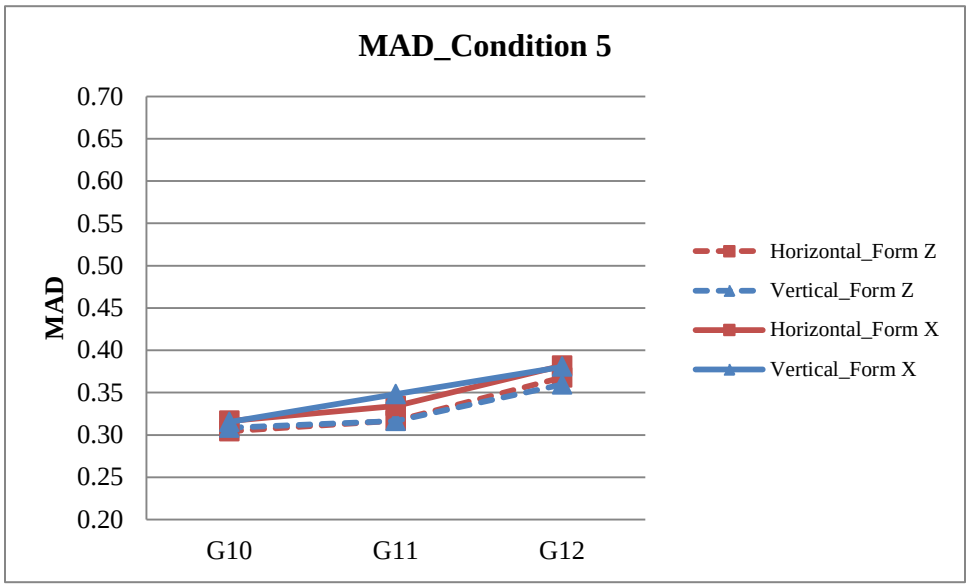


Figure 4.18 RMSEs of the Horizontal Approach and the Vertical Approach on Forms Z and X under Conditions 5 and 6



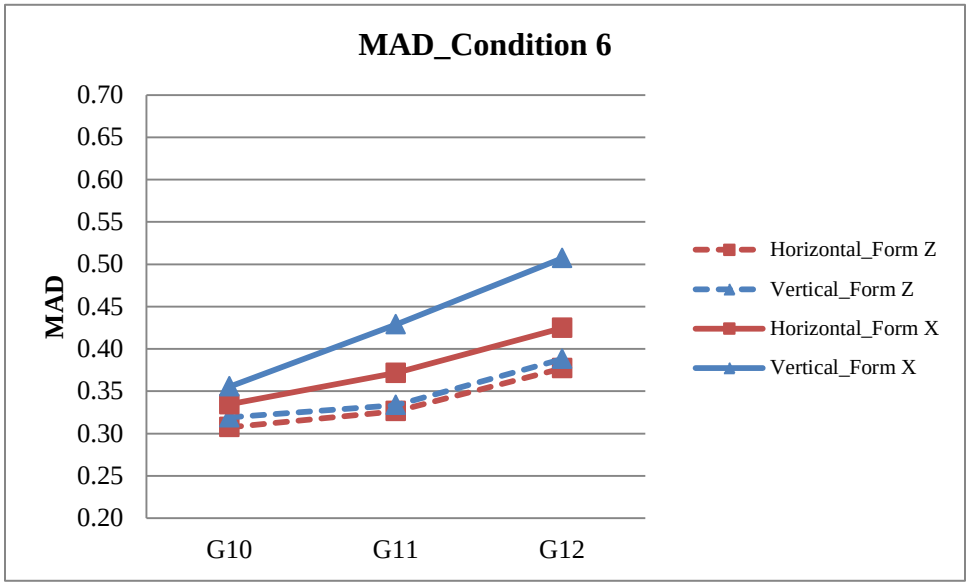


Figure 4.19 MADs of the Horizontal Approach and the Vertical Approach on Forms Z and X under Conditions 5 and 6

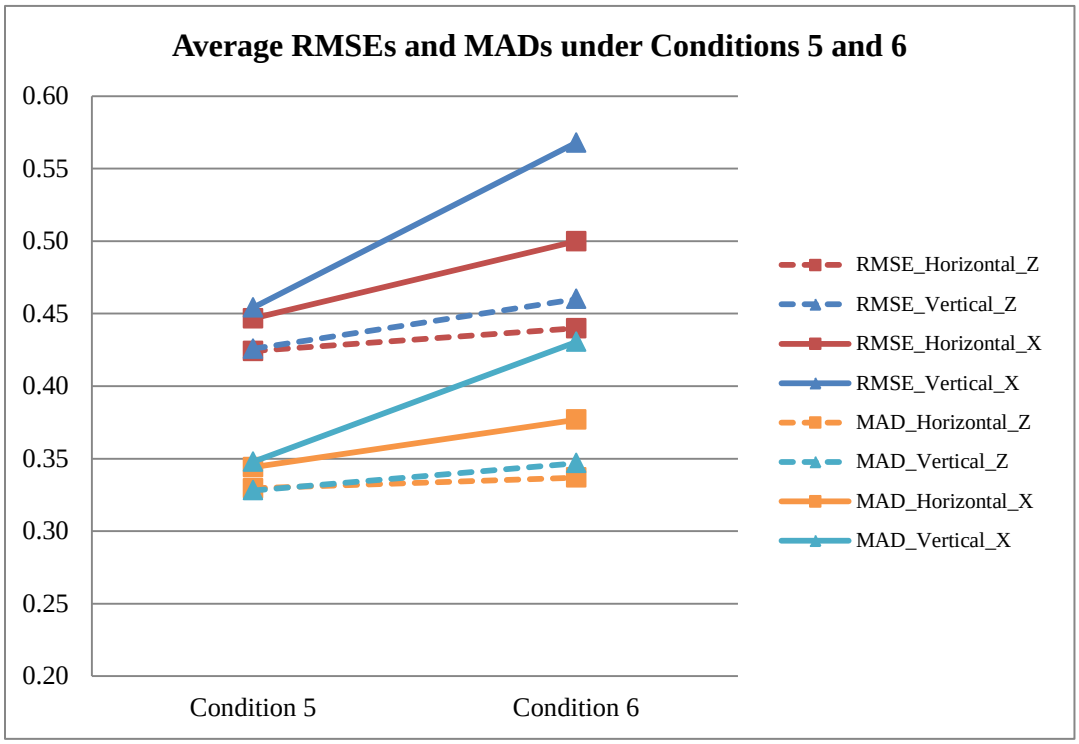


Figure 4.20 Average RMSEs and MADs under Conditions 5 and 6

Table 4.19 Grade-to-grade Means, SDs, Mean Differences, and Effect Sizes under Condition 5

	Form Z			Form X		
Condition 5	TRUE	Horizontal	Vertical	TRUE	Horizontal	Vertical
Grade			Mean			
10	0.249	0.253	0.199	0.497	0.534	0.426
11	0.919	0.887	0.893	1.167	1.082	1.019
12	1.272	1.227	1.241	1.524	1.547	1.461
Grade			SD			
10	1.001	0.963	0.970	0.998	0.881	0.935
11	1.200	1.120	1.113	1.200	1.133	1.111
12	1.401	1.313	1.199	1.400	1.300	1.281
Grade			Mean Difference			
10/11	0.670	0.634	0.694	0.670	0.549	0.593
11/12	0.353	0.340	0.348	0.357	0.465	0.441
Grade			Effect Size			
10/11	0.606	0.607	0.665	0.607	0.541	0.577
11/12	0.271	0.278	0.301	0.273	0.381	0.368

Table 4.20 Grade-to-grade Means, SDs, Mean Differences, and Effect Sizes under Condition 6

	Form Z			Form X		
Condition 6	TRUE	Horizontal	Vertical	TRUE	Horizontal	Vertical
Grade			Mean			
10	0.499	0.499	0.401	0.997	1.021	0.827
11	1.169	1.115	1.064	1.667	1.547	1.372
12	1.522	1.463	1.376	2.024	1.998	1.757
Grade			SD			
10	1.001	0.954	0.951	0.998	0.876	0.900
11	1.200	1.068	1.050	1.200	1.030	1.059
12	1.401	1.260	1.110	1.400	1.204	1.200
Grade			Mean Difference			
10/11	0.670	0.617	0.663	0.670	0.526	0.545
11/12	0.353	0.348	0.312	0.357	0.452	0.385
Grade			Effect Size			
10/11	0.606	0.609	0.661	0.607	0.550	0.555
11/12	0.271	0.298	0.289	0.273	0.403	0.340

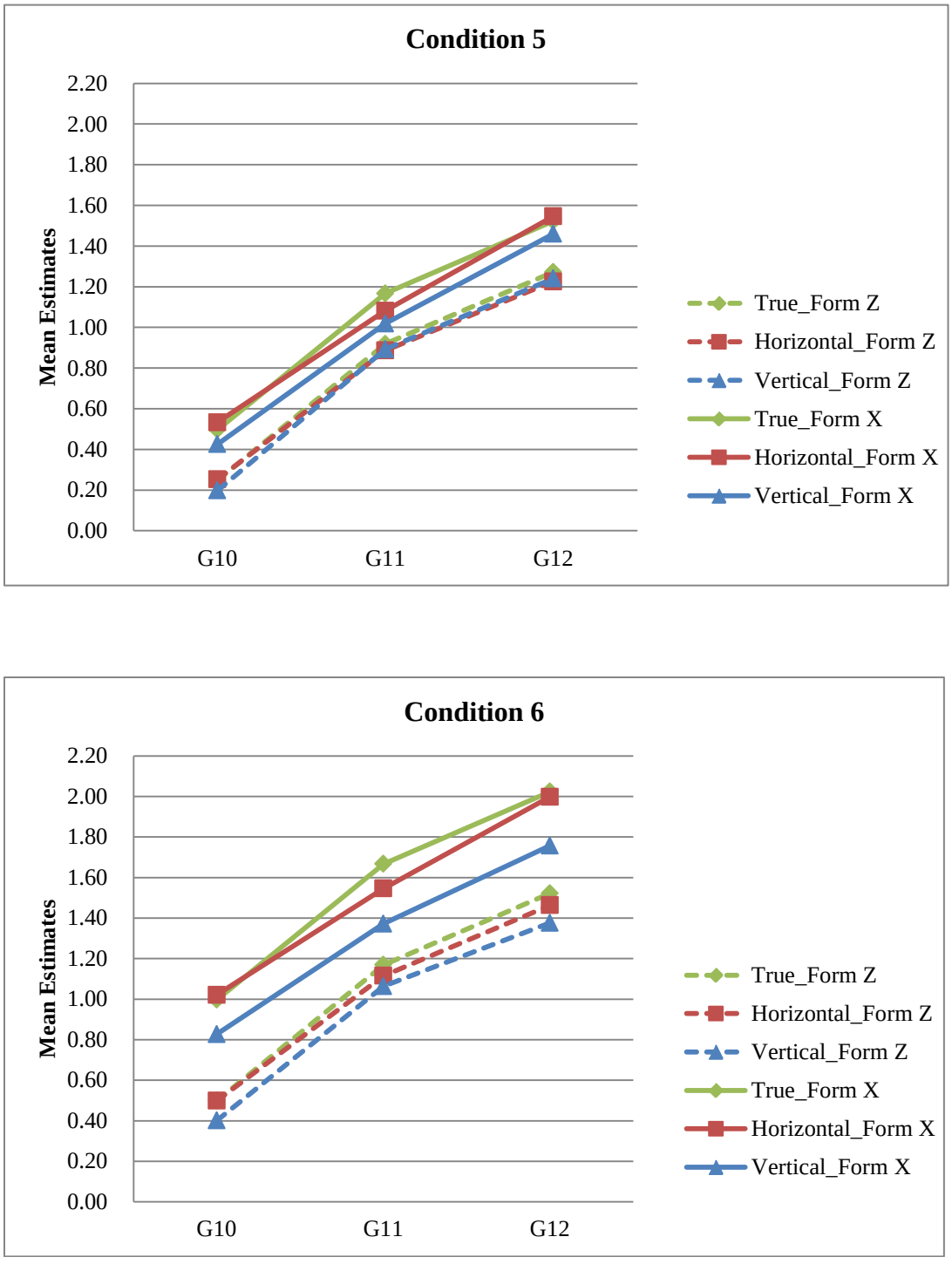
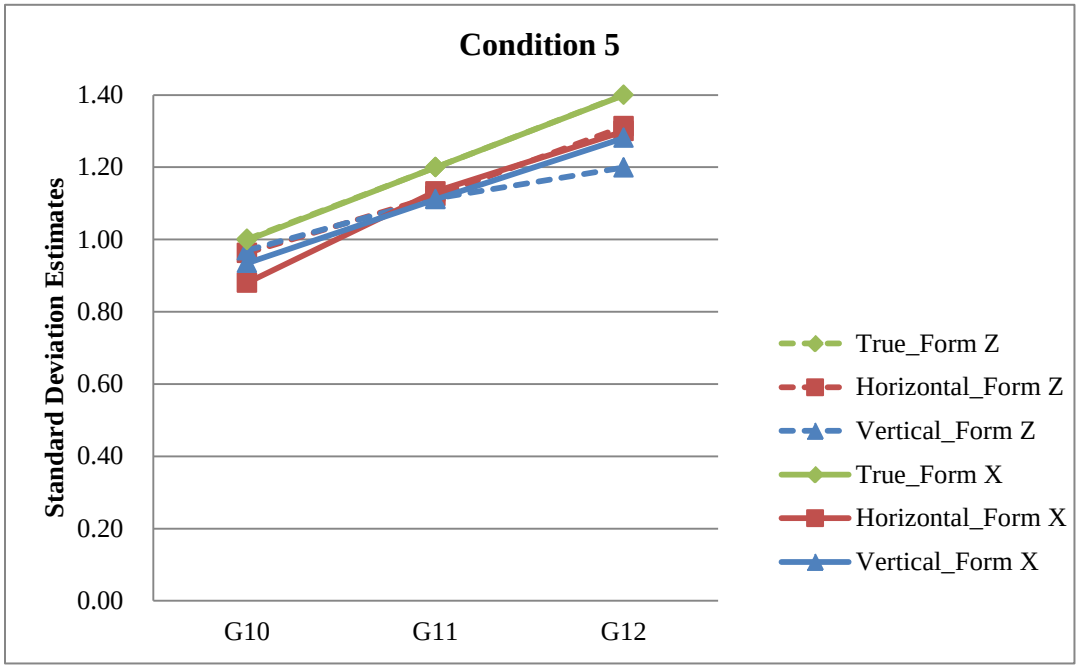


Figure 4.21 Grade-to-grade Mean Estimates under Conditions 5 and 6



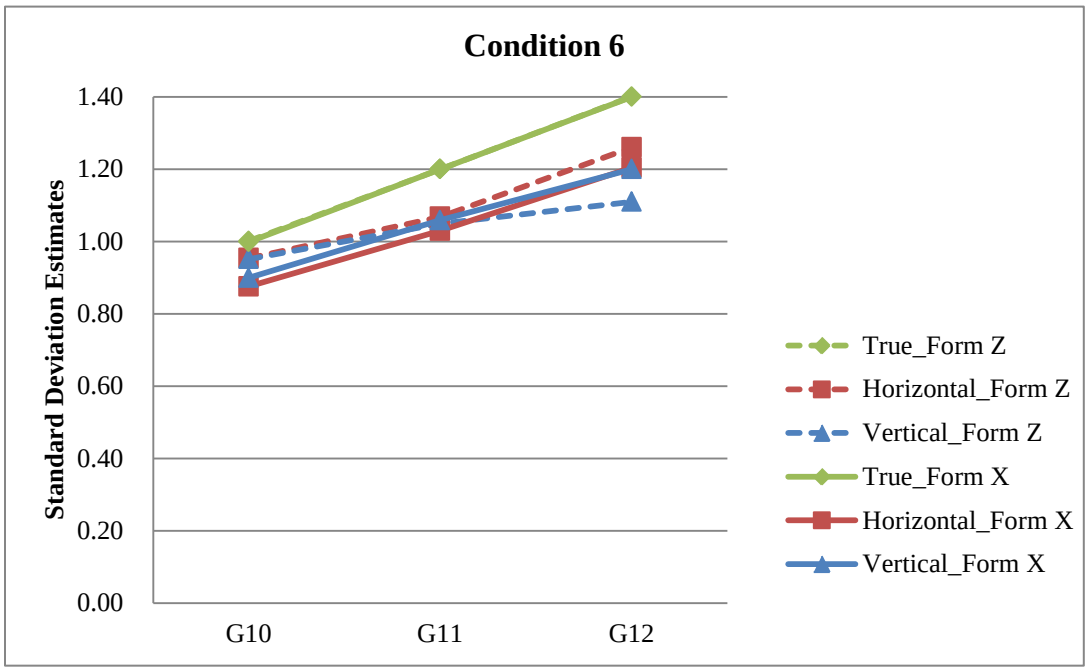


Figure 4.22 Grade-to-grade Standard Deviation Estimates of Ability Estimates under Conditions 5 and 6

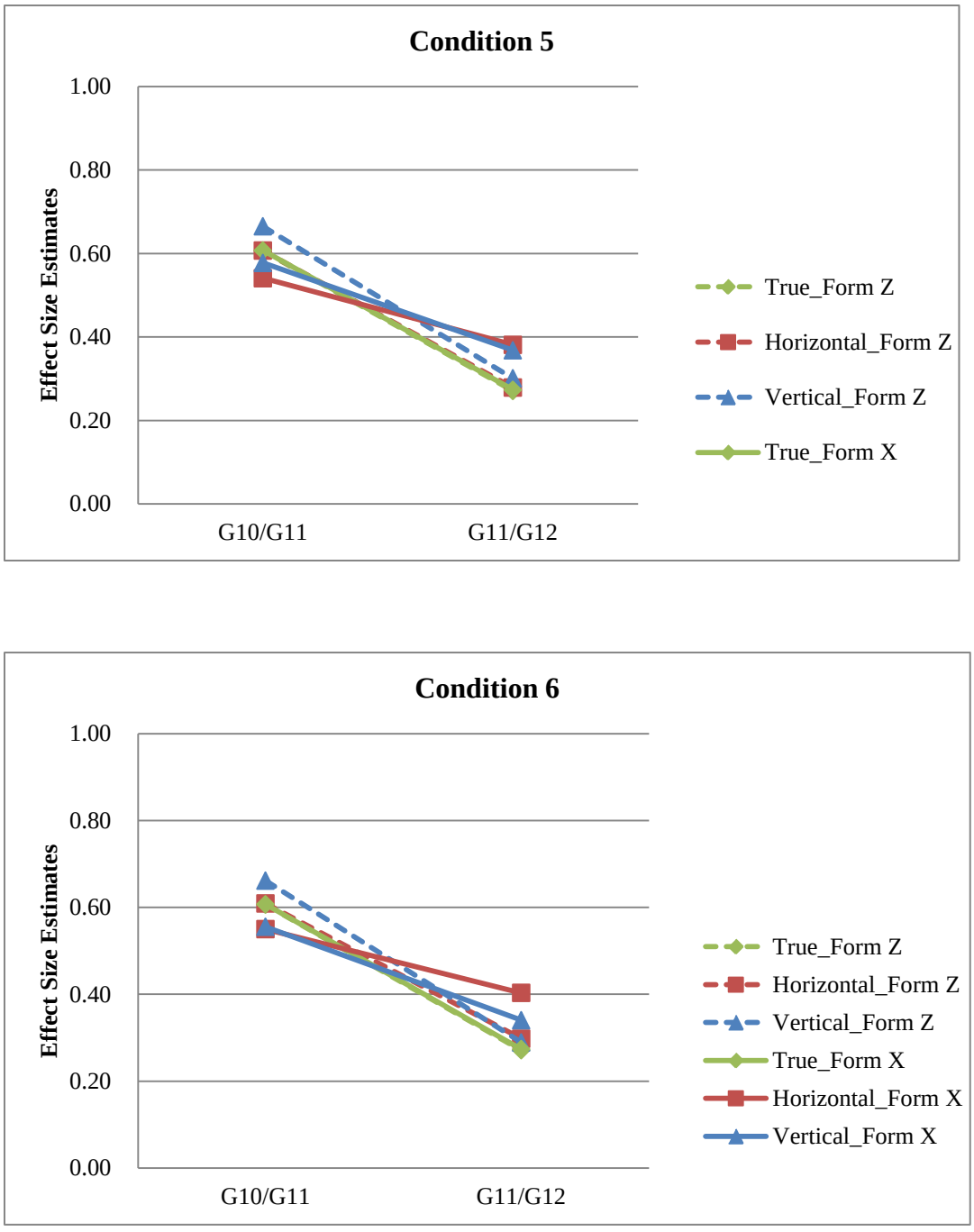
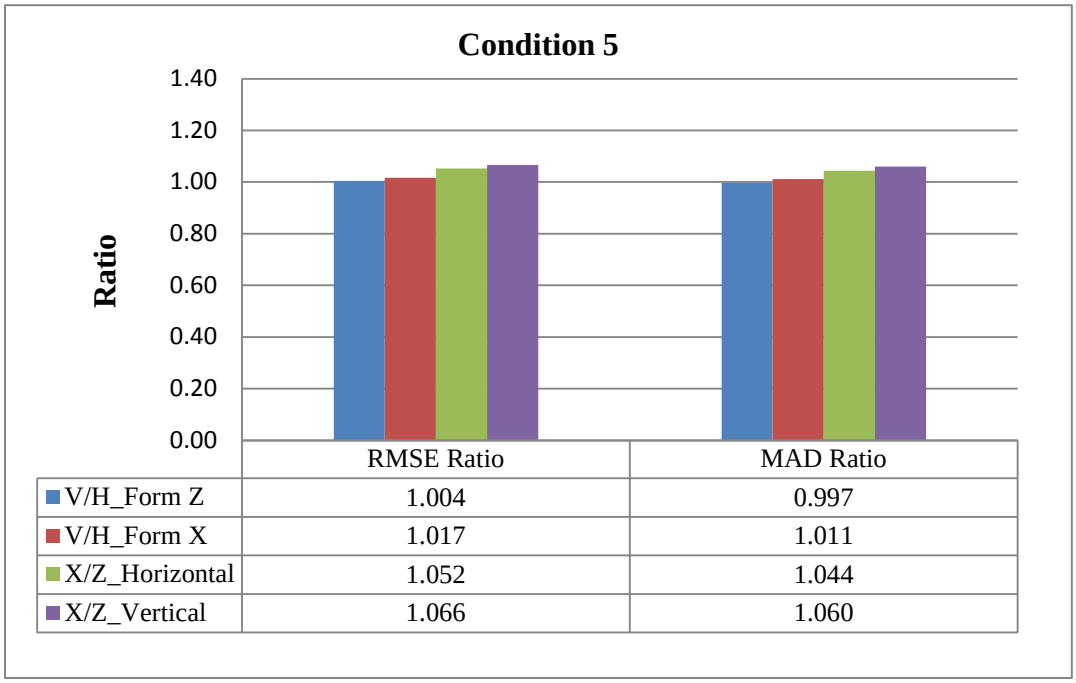
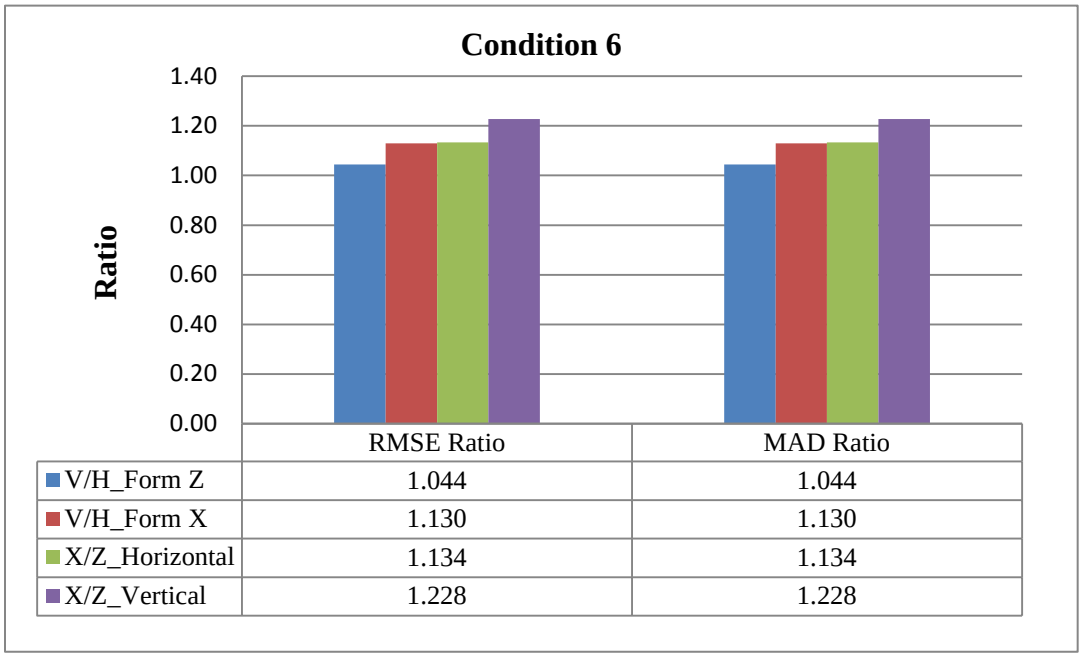


Figure 4.23 Grade-to-grade Effect Size Estimates of Ability Estimates under Conditions 5 and 6



	RMSE Ratio	MAD Ratio
V/H_Form Z	1.004	0.997
V/H_Form X	1.017	1.011
X/Z_Horizontal	1.052	1.044
X/Z_Vertical	1.066	1.060



	RMSE Ratio	MAD Ratio
V/H_Form Z	1.044	1.044
V/H_Form X	1.130	1.130
X/Z_Horizontal	1.134	1.134
X/Z_Vertical	1.228	1.228

Figure 4.24 RMSE Ratios and MAD Ratios between the Scale Maintenance Approaches and between Forms under Conditions 5 and 6

CHAPTER V

DISCUSSION AND CONCLUSION

Following *the Reauthorization of the Elementary and Secondary Education* (USDE, 2010), which requires all states to develop college- and career-ready standards and thus to prepare college- and career-ready students, monitoring students' progress from year to year and over the course of their studies is essential. Vertical scales can assess growth at the student level and provide the assessment of progress towards goals in subsequent grades on the same metric. When vertical scales are well defined and constructed for use in large-scale educational assessment programs, they can significantly enrich the interpretation of test scores by providing a systematic way to examine student knowledge and skill acquisition with respect to implemented curricula across grade spans. When new forms of the assessment are developed, an equating design must be incorporated as part of the implementation of the new forms. A form-to-form vertical scale equating design should foster scale comparability of the scale across grades within form, within grade across forms, and across grades and across forms.

The purposes of this dissertation are to compare how different the resulting proficiency estimates are by using two scale maintenance approaches, the horizontal and vertical approaches, in supporting scale stability across grade within form, within grade across forms, and across grades and across forms, and to thus investigate under which conditions of within-grade variability patterns and examinee sample characteristics one

approach is preferable to the other. The research questions are presented in the Chapter I of this dissertation. Since there is no universally accepted growth model in the literature, three different distribution sets were specified and generated with regard to within-grade variability patterns in the simulation data: constant across grades, decreasing variability as grade increases, and increasing variability as grade increases. In addition, two sets of examinee sample characteristics were also specified in the simulation data: small examinee group difference and large examinee group difference. Thus six proficiency distribution conditions were used to generate data in this dissertation. Under the six conditions of proficiency distributions, the performances of the two scale maintenance approaches on the resulting proficiency estimates across multiple forms were the foci in this dissertation.

In this chapter, the first section provides a summary and discussion of the results. Each of the research questions is revisited and the results addressing these questions are described and discussed. In the second and third sections, limitations of the study and further directions are discussed. The educational importance of this study concludes the last section.

Summary and Discussion of the Main Questions

The issue of how to maintain a vertical scale over time has been largely ignored in the literature until recent years. Almost all the research on this topic has focused on whether vertically linking one form of a multi-level form to another is preferable to maintaining a multi-level scale by horizontal equating individual levels across two forms

or over two years (see Cao, Li, & Hendrickson, 2007; Hoskens, Lewis, & Patz, 2003; Tong & Kolen, 2008; Tong & Kolen, 2009; Wang & Harris, 2009a; Wang & Harris, 2009b). However, the other studies all used empirical data except Cao, Li, and Hendrickson's study and none of them spanned three or more years. One of the previous studies did investigate the topic of maintaining a baseline vertical scale over three years, but the study also used empirical data from a state assessment program. Therefore, it is difficult to generalize the results of these studies. Two commonalities of these previous studies are that all of them used the IRT models and the common-item design. This motivated the use of similar methodology including the IRT 3PLM in this present study as well as the common-item design for data generation including both horizontal and vertical common items. By using the IRT 3PLM and the common-item nonequivalent groups design, this study adds to this body of research by comparing the performances of different scale maintenance approaches in maintaining a baseline vertical scale across three forms. No other studies were found that manipulated the conditions of within-grade variability patterns and examinee sample characteristics and evaluated how different the resulting proficiency scale conversions were on the new form by different scale maintenance approaches under these manipulations.

This study compared the resulting proficiency estimates by the two scale maintenance approaches using several common criteria found in the psychometric literature. Pearson product-moment coefficients, RMSE, and MAD are the three indices used to investigate how adequate the scale maintenance approaches are in recovering the

true examinee proficiencies under the six conditions. To investigate how different the resulting proficiency estimates are by the two scale maintenance approaches, three properties are used: patterns of grade-to-grade growth, patterns of grade-to-grade variability and separation of grade distributions. The indices -- mean, mean difference between adjacent grades, SD, and effect size -- are computed for all the estimated proficiency distributions. To evaluate under which condition one approach is better than the other, the RMSE ratios and the MAD ratios between the two maintenance approaches within form are compared. In addition, the RMSE ratios and MAD ratios by each of the two approaches are computed between the new Form X and the interim Form Z.

Research Question 1

Research Question 1 was investigated under the six conditions of within-grade variability patterns and examinee sample characteristics:

> *How adequate are the two scale maintenance approaches in recovering the true examinee proficiencies, under the six conditions of proficiency distributions, by considering the factors of within-grade variability patterns and examinee sample characteristics?*

On the interim Form Z with a single linking to the base Form Y, the horizontal approach and the vertical approach yielded similar correlations between the true proficiency and the proficiency estimate, regardless of the conditions of within-grade variability patterns and examinee group differences. This finding is consistent with the findings in Cao, Li, & Hendrickson's study (2007). Moreover, on the new Form X with

multiple linkings to the base Form Y, the values of correlations are very close between the two approaches under each of the six conditions. This indicates that, when linking across more forms or more years, the horizontal approach and the vertical approach produce similar results in recovering the true examinee proficiencies across three forms, regardless of the within-grade variability patterns and examinee group differences.

However, the magnitudes of the paired correlations using the horizontal approach and the vertical approach are different, by comparing within and among the six conditions. Specifically, under each of the six conditions, the accuracies in recovering the true examinee proficiencies by the two approaches tend to be less on the new Form X than on the new Form Z. This is not unexpected as Form X involves an additional linking. This is most noticeable under the conditions with large examinee group differences, especially when within-grade variability remains constant or decreases.

By comparing the resulting average RMSEs and MADs between the two scale maintenance approaches within form under each condition, the horizontal approach yielded slightly smaller RMSEs and MADs than the vertical approach. These results are not surprising since more transformations are involved in the vertical approach than the horizontal approach which may introduce relatively more error, but the differences are small and probably can be ignored in this situation. However, the RMSEs and MADs of both approaches increased as three forms were involved, and the increasing tendencies were more evident under the conditions of large examinee group difference. If more than

three forms were to be involved, the increasing tendencies of the RMSE and MAD may be more noticeable.

Overall, the results on adequacy in recovering the true proficiencies on the new Form X through multiple linkings suggest that the adequacy in recovering the true examinee proficiencies on the new form across multiple linkings appeared to be an interaction among the number of forms or years involved in the linking process, the within-grade variability patterns, and the examinee group differences. However the accuracy appeared not to be a function of the maintenance approach implemented.

Research Question 2

Research Question 2 was investigated under the six conditions of within-grade variability pattern and examinee sample characteristics:

> *What are the effects of the two scale maintenance approaches on the resulting proficiency estimates and growth interpretations, under the six conditions of proficiency distributions, considering the factors of within-grade variability patterns and examinee sample characteristics?*

As discussed in Chapter II, previous studies found that it is hard to decide which method is superior in capturing the underlying true scales between the horizontal approach and the vertical approach when maintaining the baseline vertical scale across two forms. The analyses of the simulated data under the six conditions in this study showed that both the horizontal approach and the vertical approach were able to capture the general characteristics of the underlying true growth trends on the baseline form Y

when single linking involved between the baseline Form Y and the interim Form Z. The results on the interim Form Z in this study confirm in the available literature on maintaining a baseline vertical scale across two forms (Cao et al., 2007; Tong & Kolen, 2008; Tong & Kolen, 2009; Wang & Harris, 2009a; Wang & Harris, 2009b). When multiple linking involves across three forms, the findings of this study provide more information beyond what has been found in the existing literature.

Grade-to-grade Growth

In terms of grade-to-grade growth, both approaches were able to capture the increasing trend of the true proficiencies as grades increased under the six conditions. Yet, differences between the two scale maintenance approaches were detected, when multiple linking involved the new form X.

With regard to the value discrepancy between the true means and the mean estimates, the horizontal approach tended to perform better than the vertical approach. This advantage of the horizontal approach was more evident on the new Form X than on the interim Form Z. However, when within-grade variability remained constant or increased with large examinee group differences, the horizontal approach tended to underestimate the true means at the higher Grade 12, and thus showed less growth at Grade 12 than the true growth. This approach, however, produced an overestimate at the higher Grade 12 with the constant within-grade variability pattern when examinee group differences across forms were small. These findings indicated that horizontal approach tended to be less affected by the number of forms to be linked, but was affected by an

interaction of the within-grade variability patterns and examinee group differences at higher Grade 12. A possible explanation for the larger extent of over/under-estimations on Form X would be that multiple linking caused more estimation discrepancies on Form X, especially at higher grade when more transformations and more links were involved through a linking chain.

The vertical approach produced underestimates of the true means when the within-grade variability remained constant or decreased, and the extent of underestimation increased as examinee group differences became larger. However, when the within-grade variability increased as grade increased, the vertical approach tended to overestimate the true means at the lower Grades 10 and 11. Though the extent of overestimation was larger on the new Form X than on the interim Form Z, the extent of overestimation appeared not to be affected by large examinee group difference. One implication of these findings is that the vertical approach is affected by an interaction of the number of forms to be linked and examinee group differences when the within-grade variability remains constant or decreases as grade increase, but when the within-grade variability increased across grades, the vertical approach appeared to be a function of number of forms to be linked but not a function of examinee group differences.

In sum, grade-to-grade growth appeared to be a function of the scale maintenance approach, the number of forms to be linked, the within-grade variability patterns, and examinee group differences. In regard to the first research question, the adequacy in recovering the true proficiency appeared to be independent from the scale maintenance

approaches implemented. However, the grade-to-grade growth appeared to be dependent on the scale maintenance approaches implemented.

Grade-to-grade Variability & Separation of Grade Distributions

To estimate the within-grade variability, both approaches were able to capture the general changing trend across the three pairs of distribution sets: constant SDs, decreasing SDs, and increasing SDs as grade increases. Both approaches tended to underestimate the true SDs under the six conditions. When the true SDs remained constant across grades, the vertical approach yielded a more parallel trend to the true constant within-grade variability pattern but with larger distances from the true one when examinee group differences were small; the horizontal approach tended to yield flatter trends when examinee group differences were large. With regard to the decreasing and increasing within-grade variability patterns, which scale maintenance approach was superior in capturing the true within-grade variability patterns was unclear. The horizontal approach and the vertical approach showed similar decreasing or increasing trends across grades.

For effect size, the results from both approaches suggested that students' growth decelerates as grade increased under all the six conditions, in agreement with the true underlying scale. Yet both approaches showed fluctuations under different conditions. In general, the vertical approach was slightly superior with respect to estimation accuracy, and this trend can be better observed at higher Grades 11 and 12 under the six conditions.

Overall, the grade-to-grade variability appeared to be function of the scale maintenance approach implemented and the examinee group differences when the within-grade variability remained constant as grades increase. The growth deceleration as grades increase appeared to be a function of the scale maintenance approach implemented and grade levels but not a function of the number of forms to be linked, the within-grade variability patterns, and the examinee group differences.

Research Question 3

The third research question investigated in this study is:

Under which conditions of proficiency distributions does one scale maintenance approach provide more equivalent resulting proficiency scales than the other, by considering the factors of within-grade variability patterns and examinee sample characteristics?

Previously available studies just used the criteria for the first two research questions in this study. In this study, more criteria were used to further investigate under which condition one approach was superior to the other, the ratios of the average RMSEs and the ratios of the average MADs between the vertical approach and the horizontal approach within form as well as the ratios between the new Form X and the interim Form Z via each of the two approaches. Generally, all the ratios were greater than 1 under the six conditions, indicating that the vertical approach introduced more error than the horizontal approach and multiple linking on the new Form X introduced more error than single linking on the interim Form Z.

However, the ratios between the two scale maintenance approaches within form were very close to 1, which indicated that the differences between the horizontal approach and the vertical approach were so small that they could likely be ignored. Furthermore, the magnitudes of the ratios between the new Form X and the interim Form Z via each approach were larger than the ratios between scale maintenance approaches within form, especially when the within-grade variability remains constant or increases with large examinee group difference.

One major implication of these findings is that the accuracy in recovering the true examinee proficiencies on the new form across multiple linkings appeared to be an interaction among the number of forms or years involved in the linking process, the within-grade variability patterns, and the examinee group differences, but they did not appear to be a function of the maintenance approach implemented, which was consistent with conclusions for the first research question.

Limitations

In this dissertation, three test forms consisting of three levels each from Grade 10 to Grade 12 were generated at the same time, for the purpose of investigating these research questions. The simulated data were assumed to be normally distributed and the means and SDs were specified in Table 3.8. The within-grade variability was assumed to increase, decrease and remain constant over grades, because no-universally accepted growth model currently existed in the literature. A total 40% increase or decrease in variability was assumed from Grade 10 to Grade 12, with constant increase or decrease

for the levels in between. In addition, the examinee sample differences across forms were assumed to be small with a mean proficiency difference of 0.25 and large with a mean proficiency difference of 0.50, because simulation studies under common-item non-equivalent groups design in the literature often used these two values of mean proficiency difference between the two populations. Normal deviates were generated using the computer software R and 3PLM was used to simulate item responses of the nine level tests across the three test forms.

This approach to simulating data works well when the sole purpose is to let all the model assumptions hold for the scale maintenance process and to observe which scale maintenance approach can better capture the true growth under those conditions. But in practice, in operational testing programs data does not meet all the model conditions. Some ways to generate more realistic data should be explored, where the underlying scale is still known, but the assumption of the scale maintenance approaches do not hold exactly. In addition, data to be generated could be passage-based or include both multiple-choice items and constructed response items. Linking vertical scales across multiple forms for simulated data using the two approaches as described in this dissertation and compared with the true scale can help find the calibration and linking process that are most robust to violation of assumptions.

This study used RMSE and MAD to evaluate the accuracy of the two scale maintenance approaches in recovering the true proficiencies. Comparisons of the two indices were made between the results of the two scale maintenance approach, which

only provided information about which approach yielded relatively less error than the other under the six conditions. Given that the average correlations were in the range of 0.905 to 0.938 between the true proficiencies and the proficiency estimates on the two forms (Form Z and Form X) via the two scale maintenance approaches under the six conditions, up to 20% of the estimation has not been recovered in scale maintenance process. If more estimation had been recovered, the magnitudes of the RMSE and MAD would have been lower than what were obtained in this study.

One of the limitations of this study is that vertical scales were developed and maintained for only three grades, Grades 10-12. This would affect the extent to which generalizations can be made for more grades, such as Grades 3-12, regarding the different scale maintenance approaches under various conditions. Future research involving more grade levels could draw more generalizable conclusions on the impact of scale maintenance approaches, within-grade variability patterns, and examinee sample characteristics. More grades may lead to larger violations of unidimensionality assumption of measuring a common construct across grade levels.

Another limitation of this study is that the common items used in horizontal equating and in vertical linking were assumed to be discrete multiple-choice items. For passage-based tests, the results from this study might not be generalizable to some extent. Issues of IRT model fit and uni-/multi-dimensionality need to be further investigated. In addition, it would be more difficult to select overlapping items both for the horizontal equating between forms and for the vertical scaling between adjacent grades, which was

assumed to be a miniature set of the whole test. Future research on the same topic as this study could be conducted, by assuming that tests are passage-based. These results could then be compared to further investigate the performances of the two scale maintenance approaches with respect to different within-grade variability patterns and small/large examinee sample differences across multiple forms.

Under the data collection design of this study, both horizontal common items and vertical common items are required in the level tests, even though the vertical common items are not used in the maintenance process for the horizontal approach. However, in practice when the horizontal approach is chosen to maintain the baseline vertical scale across forms, the new forms to be linked don't include vertical common items. The findings of this study about the horizontal approach may not be universal in such empirical situations. It would be interesting to examine whether the horizontal approach would introduce less error through linking across three forms than what has been found in this study, by excluding unused vertical common items from the new forms.

In addition, the tests developed in this study were in English Language Arts only, while previously available research on how to maintain a baseline vertical scale across two forms, such as Tong and Kolen's (2009) and Tomkowicz et al.'s (2010), used empirical data in English Language Arts (ELA) and Mathematics. Since English Language Arts and Mathematics are typically considered continuous, with overlapping constructs at least adjacent grade level, it is speculated that using data either in ELA or in Mathematics won't result in much different results on the same topic as this study.

However, it has been demonstrated that the greatest changes across grades in the measurement construct occur in subject areas as social studies or science (Tomkowicz et al., 2010). It would be interesting to examine in further studies whether content areas have different impacts in the process of scale maintenance across forms by using empirical data from different subject areas.

Further Directions

In this simulation study, small differences between the two scale maintenance approaches were detected across three forms, using the 3PLM IRT model under the common-item test design. The findings of this study suggest that multiple linkings introduces more error in the process of maintaining scales across three forms and larger examinee group differences across forms tend to have more impact on the resulting proficiency estimates, especially with the vertical approach. If more error were introduced in the linking process, the following may have contributed to the differences between the two scale maintenance approaches.

One possibility was that there might be some ‘genetic’ multidimensionality with the multilevel tests, which affected the scale maintenance approach with more grade linkages. The constructs measured at lower grades may be different from those measured at higher grades (Kolen & Brennan, 2004). To address this possibility, a multidimensional IRT model could be fit to the real data for developing the tests, and the resulting item parameter estimates and the degree of multidimensionality could be used as parameters in the simulation process. Similar procedures could be used for constructing and

maintaining vertical scales, and similar statistics could be computed on the resulting scales by the two scale maintenance approaches and compared to those from the true unidimensional baseline scale.

In this study, separate calibration was used to obtain the parameter estimates for each level test and then the Stocking-Lord method was used to link all the grades together. Another possibility to further examine the different performances of the two scale maintenance approach would be to calibrate the level tests concurrently to obtain the parameter estimates, and then to conduct the linking in a similar fashion as was used in this study. In addition, different decisions on the baseline grade in constructing a vertical scale might also have contributed to the differences between the two scale maintenance approaches. This study used the lowest Grade 10 as the baseline grade. Linking the Grade 12 proficiency estimate to the baseline Grade10 scale included two steps in the linking chain. If using mid-level Grade 11 as the baseline grade, linking Grade 12 to the baseline Grade 11 would only involve one step in the linking chain. This might lead to less error introduced in constructing the baseline vertical scale and thus might have a different impact on the resulting proficiency estimates from the two scale maintenance approaches across multiple forms.

Another future direction would be to include studies of content impact. From an empirical perspective, content experts like to select off-grade level linking items according to specifications that take into consideration item content coverage of a test blueprint, item difficulty, and item content appropriateness for specific higher or lower

grade levels. Using linking items only from the lower adjacent grade, or only from higher adjacent grade, or from both lower and higher adjacent grades would likely produce different linking results. In addition, constructed response items or passage-based items in the linking item sets might also result in different parameter estimates for operational assessments on a vertical scale. In future studies, if possible, certain hypotheses need to be made about what happens empirically in linking item selections, by considering the factors mentioned above.

Educational Importance and Conclusions

The research conducted to date described in Chapter II provides little guidance as to what methods and procedures are best for maintaining vertical scales across forms or across years. This study provides some much needed analyses and longitudinal results in terms of the effect of different maintenance approaches. Under the six conditions with regard to within-grade variability patterns and examinee sample characteristics, the two scale maintenance approaches investigated in this dissertation– the horizontal approach and the vertical approach– resulted in proficiency estimates on future forms with different characteristics as measured by adequacy in recovering the true proficiency, grade-to-grade growth, grade-to-grade variability, and separation of grade distributions. The two scale maintenance approaches, even though they have not produced dramatically different looking scales across three forms, actually will lead to somewhat different proficiency estimates.

The findings from this study provide important empirical guidance to practitioners on how the vertical scale can be maintained, once a vertical scale is established. If the desired output of a maintained scale is to continue to capture the characteristics of the established scale in terms of grade separation, within-grade variability, and growth implications, the vertical approach appears to be marginally better in achieving these goals. However, the differences observed across three forms are marginal, and in some cases neither approach demonstrates superiority in preserving the same patterns at the baseline scale. Both approaches are able to reasonably well capture the trend of the baseline scale, at least across three forms.

The horizontal approach is more straight-forward and is easier to implement in practice. It requires vertical scaling items only for the base form or the first year. For future years, linking will be conducted to equate test forms and to maintain the established vertical scale from the baseline scale. The vertical approach is more complicated, demanding vertical linking items be administered in multiple years and multiple vertical scales be established on the future forms before implementing scale maintenance approaches. Due to the marginal differences observed in this study across three forms, as well as practical concerns, it appears that horizontal approach can be a good choice for the maintenance of vertical scales.

Since small differences are found in this study between the two scale maintenance approaches across three forms, it is hard to determine how many scale maintenance cycles may lead to significant differences between the two scale maintenance approaches

investigated in this study. However, one major implication of this study is that it is not sufficient to conduct scale maintenance study for only one year or just between two forms as it takes at least two scale maintenance cycles in order to detect differences between the two scale maintenance approaches. The differences not only continue but also have the tendency to become more pronounced when more maintenance cycles are involved. Therefore, if the horizontal approach is applied, it is recommended that practitioners periodically examine the stability of the proficiency estimates to ensure the estimates do not drift too much. Scale maintenance approach selection is a practical decision as well as a psychometric one. When the horizontal approach and the vertical approach do not differ much, a practitioner may go with the easier approach to implement.

Maintaining a vertical scale across multiple forms or across years is a complicated process that involves many decisions. As the previously available research and the results of this dissertation demonstrated, different choices in the process of constructing a baseline vertical scale and in the process of maintaining the established baseline scale can lead to disparate interpretations of students' growth from year to year. Given the findings from this study, a variety of different scale maintenance approaches, different within-grade variability patterns, and examinee sample differences should be investigated and compared before finalizing the scale maintenance design from both content and psychometric perspectives. If practitioners know what their data looks like and which conditions fit their data, investigations can be conducted under the conditions most likely with their data. The findings in this study can be used as guidance for their empirical

investigations. While there is much research still needed in the area of maintaining vertical scales, this dissertation adds significantly to the literature by investigating different scale maintenance approaches in linking a vertical scale across multiple forms under various conditions of within-grade variability patterns and examinee sample characteristics.

REFERENCES

Angoff, W.H., & Cowell, W.R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. *Journal of Educational Measurement, 23*, 327-345.

Baker, F.B., & Kim, S.,-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marccel Dekker, Inc.

Becker, D.F., & Forsyth, R.A. (1992). An empirical investigation of Thurstone and IRT methods of scaling achievement test. *Journal of Educational Measurement, 29*, 341-354.

Beguin, A.A., & Hanson, B.A. (2001, April). *Effect of non-compensatory multidimentionality on separate and concurrent estimation in IRT observed score equating.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle. WA.

Beguin, A.A., Hanson, B.A., & Glas, C.A.W. (2000, April). *Effect of multidimentionality on separate and concurrent estimation in IRT equating.* Paper presented at the Annual Meeting of the American Educational Research Association, Seattle. WA.

Bock, R.D. (1997). The nominal categories model. In W.J. van der Linden & R.K. Hambleton (Eds.). *Handbook of modern item response theory* (pp. 34-39). New York: Springer-Verlag.

Bock, R.D., & Zimowski, M.F. (1997). Multiple group IRT. In W.J. van der Linden & R.K. Hambleton (Eds.). *Handbook of modern item response theory* (pp.433-448). New York: Springer-Verlag.

Briggs, D.C., Weeks, J. P., & Wiley, E. (2008). *The impact of vertical scaling decisions on growth projections.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York City.

Cao, Y., Li, D.Y., & Hendrickson, A.B. (2007). *Maintaining vertical scale under the common-item design: A simulation Study.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Camilli, G.(1988). Measurement error, multidimensionality, and scale shrinkage: A reply to Yen and Burket. *Journal of Educational Measurement, 36*(1), 73-78.

Camilli, G., Yamamoto, K., & Wang, M.-M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17*(4), 379-388.

Chin, T.Y., Kim, W., & Nering, M.L. (2006). *Five statistical factors that influence IRT vertical scaling*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Goulet, L.R., Linn, R.L., & Tatsuoka, M.M. (1975). *Investigation of methodological problems in educational research: Longitudinal methodology* (Project No. 4-1114). Urbana-Champaign, IL: University of Illinois at Urbana-Champaign (ERIC Document Reproduction Service No. ED124541).

Guskey, T.R. (1981). Comparison of a Rasch model scale and the grade-equivalent scale for vertical equating of test scores. *Applied Psychological Measurement, 5*(2), 187-201.

Gustafsson, J.-E. (1979). The Rasch model in vertical equating of test: A critique of Slinde and Linn. *Journal of Educational Measurement, 16*(3), 153-157.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares methods. *Japanese Psychological Research, 22*, 144-149.

Hanson, B.A., & Beguin, A.A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3-24.

Hanson, B. A., & Zeng, L. (2004). *ST: A computer program for IRT scale transformation*. Iowa City, IA: ACT.

Harris, D.J. (1987). *Estimating examinee achievement using a customized test.* Paper presented at the Annual Meeting of the American Educational Research Association, Washington D.C.

Harris, D.J. (1991). A comparison of Angoff's Design I and Design II for vertical equating using traditional and IRT methodology. *Journal of Educational Measurement, 28*(3), 221-235.

Harris, D.J. (1993). *Practical issues in equating.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta.

Harris, D.J. (2007). Practical issues in vertical scaling. In N.J. Dorans, M. Pommerich, & P. M. Holland (Eds.), *Linking and aligning scores and scales*. (pp. 233-251) New York, NY: Springer Science+Business Media, LLC.

Harris, D.J., Hendrickson, A.B., Tong, Y., Shin, S.-H., & Shyu, C-Y.(2004). *Vertical scales and the measurement of growth*. Paper presented as the Annual Meeting of the American Educational Research Association, San Diego, CA.

Harris, D.J., & Hoover, H.D. (1987). An application of the three-parameter IRT model to vertical equating. *Applied Psychological Measurement, 11*(2), 151-159.

Hendrickson, A.B., Kolen, M.J., & Tong, Y. (2004).*Comparison of IRT vertical scaling from scaling-test and common-item designs.* Paper presented as the Annual Meeting of the American Educational Research Association, San Diego, CA.

Holmes, S.E. (1982). Unidimesionality and vertical equating with the Rasch model. *Journal of Educational Measurement, 19*(2), 139-147.

Hoskens, M., Lewis, D. M., & Patz, R. J. (2003). *Maintaining vertical scales using common item design*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Hung, P., Wu, Y., & Chen, Y. (1991). *IRT item parameter linking: Relevant issues for the purpose of item banking*. Paper presented at the International Academic symposium on Psychological Measurement, Tainan, Taiwan.

Jodoin, M.G., Keller, L.A., & Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *The Journal of Experimental Education, 71*(3), 229-250.

Karkee T., Lewis, D.M., Hoskens, M., Yao, L., & Haug, C. (2003). *Separate versus concurrent calibration methods in vertical scaling.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Kim, S.-H., & Cohen, A.S. (1992).Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29*(1), 51-66.

Kim, S.-H., & Cohen, A.S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22*(2), 131-143.

Kim, J., Frisbie, D., & Kim, D.I. (2007). *A comparison of calibration methods and proficiency estimators for creating IRT vertical scales.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Kim, S., & Kolen, M.J. (2007). Effects on scale linking of different definitions of criterion functions for the IRT characteristic curve methods. *Journal of Educational and Behavioral Statistics*, *32*(4), 371-397.

Kim, S., & Lee, W. (2004). *IRT scale linking methods for mixed-format tests* (ACT Research Report 2004-5). Iowa City, IA: ACT, Inc.

Kim, J., Lee, W.C., Kim, D.I., & Kelly, K. (2009). *Investigation of vertical scaling using the Rasch model*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement*, *18* (1), 1-11.

Kolen, M. J. (2003). *Equating and vertical scaling: Research questions*. Paper presented at the Annual Meeting of the National Council on Measurement in education, Chicago. IL.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Levine, R. (1955). *Equating the score scales of alternative forms administered to samples of different ability* (Research Bulletin 55-23). Princeton, NJ: Educational Testing Service.

Lin, P., & Dorans, N.J. (2011). *Assessing population invariance of vertical linking functions*. Paper presented at the Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, LA.

Lord, F.M. (1977). A study of item bias using item characteristic curve theory. In Y.H. Poortinga (Ed.), *Basic problems in cross-cultural psychology*. Amsterdam: Swets&Zeitlinger B.V.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Loyd, B.H., & Hoover, H.D. (1980).Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 179-193.

Loyd, B.H. & Plake, B.S. (1987). *Vertical equating: Effects of model, method, and content domain*. Paper presented at the Annual Meeting of American Educational Research Association, Washington D. C.

Macro, G.L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139-160.

Macro, G.L., Petersen, N.S., & Stewart, E.E. (1979). *A test of the adequacy of curvilinear score equating models.* Paper presented at the Computerized Adaptive Testing Conference, Minneapolis, MN.

Macro, G.L., Peterson, N.S., & Stewart, E.E. (1983). A test of the adequacy of curvilinear score equating models. In D. Weiss (Ed.), *New horizons in testing* (pp.147-176). New York: Academic.

Meng, H., Wang, T., Vispoel, W., Lee, W.C., & Wang, C. (2007). *A comparison of IRT calibration methods for mixed-format tests in vertical scaling.* Paper presented at the Annual Meeting of the National Council on Measurement in education, Chicago. IL.

Meng, H., Kolen, M.J. & Lohman, D. (2006, April). *An empirical investigation of IRT scaling methods: How different IRT models, parameter estimation procedures, proficiency estimation methods, and estimation programs affect the results of vertical scaling for the Cognitive Abilities Test.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Muraki, E. (1997). A generalized partial credit model. In W.J. van der Linden & R.K. Hambleton (Eds.). *Handbook of modern item response theory* (pp. 153-164). New York: Springer-Verlag.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Nozawa, Y. (2008). *Comparison of parametric and nonparametric IRT equating methods under common-item nonequivalent groups design*. Unpublished Ph.D. dissertation, The University of Iowa.

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Economic Review, Otaru University of Commerce, 51*(1), 1-23

Ogasawara, H. (2001b). Least squares estimation of item response theory linking coefficients. A*pplied Psychological Measurement, 25*(4), 3-24.

Ogasawara, H. (2001c). Marginal maximum likelihood estimation of item response theory (IRT) equating coefficients for the common-examinee design. *Japanese Psychological Research, 43*(2), 72-82.

Patience, W.M. (1981). *A comparison of latent trait and equipercentile methods of vertical equating tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Los Angeles, CA.

Patz, R.J., & Hanson, B.A. (2002). *Psychometric issues in vertical scaling*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Pearson, K. (1896). Mathematical contributions to the theory of evolution – III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society of London*, series A, 187, 253-318.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989).Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221-262). New York: Academic Press, Inc.

Phillips, S.E. (1983). Comparison of equipercentile and item response theory equating when the scaling test method is applied to a multilevel achievement battery. *Applied Psychological Measurement, 7*(3), 267-281.

Power, S., Turhan, A., & Binici, S. (2012). *Population invariance of vertical scaling results.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

Proctor, T.P. (2008). *An investigation of effects of varying the domain definition of science and method of scaling on a vertical scale.* Unpublished Ph.D. Dissertation, The University of Iowa.

Rentz, R.R., & Bashaw, W.L. (1971). The national Reference Scale for reading: An application of the Rasch model. *Journal of Educational Measurement, 14*, 161-179.

Samejima, F. (1997). Graded response model. In W.J. van der Linden & R.K. Hambleton (Eds.). *Handbook of modern item response theory* (pp. 85-100). New York: Springer-Verlag.

Schulz, E.M, Perlman, C., Rice, W.K., Jr.,& Wright, B.D. (1992). Vertically equating reading tests*:* An example from Chicago Public Schools. In M. Wilson (Ed.) *Objective measurement: Theory into practice* (pp. 65-83), Vol. J.

Skagg, G., & Lissitz, R.W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56,* 495-529.

Skaggs, G., & Lissitz, R.W. (1988).Effect of examinee ability on test equating invariance. *Applied Psychological Measurement, 14*(1*),* 23-32.

Skaggs, G. (1990a). *Assessing the utility of item response theory models for test equating*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston, MA.

Skaggs, G. (1990b). To match or not to match samples on ability for equating: A discussion of five articles. *Applied Measurement in Education, 3*(1), 105-113.

Slinde, J.A., & Linn, R.L. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. *Journal of Educational Measurement, 15*(1), 23-35.

Slinde, J.A., & Linn, R.L. (1979). A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. *Journal of Educational Measurement, 16* (3), 159-165.

Spelling, M. (2005). *Letter to Chief State School Officers regarding a pilot project on growth.* From http:/www.ed.gov/policy/elsec/guid/secletter/051121.html.

Stocking, M.L., & Eignor, D.R. (1986). *The impact of different ability distributions on IRT pre-equating* (Research Report 86-49). Princeton, NJ: Educational Testing Service.

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

Thissen, D., Chen, W-H, & Bock, R.D. (2003). *Multilog (version 7)* [Computer software]. In Mathidal du Toit (Eds.).*IRT From SSI: BILOG-MG MULTILOG PARSCALE TESTFACT*. Chicago: Scientific Software International.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test Scoring*. (pp. 73-140). Mahwah, NJ: Lawrence Erlbaum Associates.

Thurstone, L.L. (1925). A method of scaling psychological and educational tests. *The Journal of Educational Psychology*, *16*(7), 433-451.

Thurstone, L.L. (1938). Primary mental abilities. *Psychometric Monographs.No.1*.

Tinsley, H.E., & Dawis, R.V. (1975). An investigation of the Rasch simple logistic model: Sample-free items and test calibration. *Educational and Psychological Measurement*, *35*, 325-339.

Topczewski, A. (2010). *Effect of Calibration on Vertical Scaling*. Paper presented at the Annual Meeting of the National Council on Measurement in Education. Vancouver, British Columbia, Canada.

Tong, Y. (2005).Comparisons of methodologies and results in vertical scaling for educational achievement tests. [Doctoral dissertation]. *Dissertation Abstracts International*, 66(04), 1334A. (UMI No. 9315947).

Tong, Y., & Kolen, M.J. (2007).Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, *20*, 227-253.

Tong, Y., & Kolen, M.J. (2008). *Maintenance of vertical scales*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York City.

Tong, Y., & Kolen, M.J. (2009). *A further look into the maintenance of vertical scale*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

United States Department of Education (2006). *Peer review guidance for the NCLB growth model pilot applications.* From http:/www.ed.gov/elsec/guide/ growthmodelingguidance.doc.

United States Department of Education (2010). *ESEA Blueprint for Reform.* Office of Planning, Evaluation and Policy Development, Washington, D.C.

Wang, T., Kolen, M.J., & Harris, D.J. (2000). Psychometric properties of scale scores and performance levels for performance assessment using polytomous IRT. *Journal of Educational Measurement*, *37*, 141-162.

Wang, T. & Brennan, R.L. (2009). A modified frequency estimation equating method for the common-item nonequivalent groups design. *Applied Psychological Measurement, 33,* 118-132.

Wang, T., Lee, W., Brennan, R.L., & Kolen, M.J. (In press). A comparison of the frequency estimation and chained equipercentile methods under the common-item non-equivalent groups design. *Applied Psychological Measurement*.

Wang, X., & Harris, D.J. (2009a).*Maintenance of vertical scales using simulated data based on different across-grade patterns of proficiency scale variability.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Wang, X., & Harris, D.J. (2009b). *Maintenance of IRT-based vertical scales under common-item nonequivalent groups design.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA

Way, W.D., & Tang, K.L. (1991). *A comparison of four logistic model equating methods*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Whitely, S.E., & Dawis, R.V. (1974). The nature of objectivity with the Rasch model. *Journal of Educational Measurement, 11,* 163-178.

Williams, V.S.L., Pommerich, M., &Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement, 35*(2), 93-107.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23,* 299-325.

Yen, W. M., & Bucket, G. R. (1997). Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement, 34*(4), 293-313.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items.* Chicago: Scientific Software International Inc.

APPENDIX

BILOG-MG 3 CODES

```
PROJECT
Fitting the 3PL Model with Bilog-MG 3 FOR Form Y (N=2000 n=40)
>GLOBAL  DFNAME='C:\Users\Desktop\dat.txt',
      NPARM=3,SAVE;
>SAVE    PARM='C:\Users\Desktop\PARM.txt',
      SCORE='C:\Users\Desktop\SCORE.txt';
>LENGTH  NITEMS=40;
>INPUT   TYPE=1,NIDCHAR=4;
>ITEMS   INUMBERS=(1(1)40);
>TEST    INUMBERS=(1(1)40);
(4A1,1X,40A1)
>CALIB   NQPT=40,FLOAT;
>SCORE   NQPT=40,INFO=2;
```